Data-Driven Contextual Optimization with Gaussian Mixtures: Flow-Based Generalization, Robust Models, and Multistage Extensions

YoungChul Yoon*, Grani A. Hanasusanto*, and Yijie Wang†

Contextual optimization enhances decision quality by leveraging side information to improve predictions of uncertain parameters. However, existing approaches face significant challenges when dealing with multimodal or mixtures of distributions. The inherent complexity of such structures often precludes an explicit functional relationship between the contextual information and the uncertain parameters, limiting the direct applicability of parametric models. Conversely, while non-parametric models offer greater representational flexibility, they are plagued by the "curse of dimensionality," leading to unsatisfactory performance in high-dimensional problems. To address these challenges, this paper proposes a novel contextual optimization framework based on Gaussian Mixture Models (GMMs). This model naturally bridges the gap between parametric and non-parametric approaches, inheriting the favorable sample complexity of parametric models while retaining the expressiveness of non-parametric schemes. By employing normalizing flows, we further relax the GM assumption and extend our framework to arbitrary distributions. Finally, inspired by the structural properties of GMMs, we design a novel GMM-based solution scheme for multistage stochastic optimization problems with Markovian uncertainty. This method exhibits significantly better sample complexity compared to traditional approaches, offering a powerful methodology for solving long-horizon, high-dimensional multistage problems. We demonstrate the effectiveness of our framework through extensive numerical experiments on a series of operations management problems. The results show that our proposed approach consistently outperforms state-of-the-art methods, underscoring its practical value for complex decision-making problems under uncertainty.

Key words: contextual stochastic optimization, Gaussian mixture models, distributionally robust optimization, normalizing flows, multistage stochastic programming, data-driven dynamic programming

1. Introduction

Business decision-making and operations management are fundamentally intertwined with the presence of uncertainty. Across different industries, managers are constantly tasked with making critical choices whose outcomes depend on future events that cannot be perfectly predicted. Therefore, developing solution schemes that can effectively model and respond to such uncertainty is a critical research and practical endeavor. Classical stochastic optimization approaches solely focus

* Department of Industrial and Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. Email: ycyoon2, gah@illinois.edu.

† School of Economics and Management, Tongji University, Shanghai, China. Email: yijiewang@tongji.edu.cn.

on the probability distributions of the underlying random variables. A standard formulation of a stochastic optimization problem can be expressed as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right],$$

where $\boldsymbol{x}$ represents the decision variables, $\mathcal{X}$ denotes its feasible region, and $\mathbb{P}$ characterizes the distribution of the random variable $\tilde{\boldsymbol{\xi}}$. While stochastic optimization provides a powerful framework for solving decision-making problems under uncertainty, this approach often overlooks valuable information that could provide a more accurate description of the uncertain problem. In many practical settings, decision-makers have access to observable exogenous factors, often referred to as contextual information or side information, that carry significant predictive power over the uncertain outcomes. For instance, in inventory management, future demand for certain products is often correlated with observable covariates like past sales history, promotional activities, or even macroeconomic trends (Chen and Plambeck 2008, Zhang et al. 2020). Similarly, in energy systems, the electricity generation from renewable sources like wind or solar, as well as its demand, are heavily dependent on weather conditions and seasonal patterns (Conejo et al. 2005, Ward 2013, Bhatti and Danilovic 2018). In revenue management, customer characteristics, browsing behavior on the e-commerce platform, and competitors' prices can provide valuable context for predicting willingness-to-pay and improving pricing strategies (Chen et al. 2022). Furthermore, within financial applications like asset management, stock returns are known to be influenced by a wide array of macroeconomic indicators such as interest rates, inflation rates, and GDP growth, as well as firm-specific data like market capitalization and book-to-market ratios (Fama and French 2015, Gu et al. 2020, Leippold et al. 2022).

Leveraging this contextual information allows for a more refined understanding of the underlying uncertainties, reflecting its conditional distribution rather than just its overall average behavior. This class of problem is known as contextual optimization. In this setting, prior to making a decision, the decision-maker observes a realization $\boldsymbol{s}$ of a vector of random exogenous covariates $\tilde{\boldsymbol{s}}$. Subsequently, the objective is to minimize the expected loss conditioned on the observed contextual information $\tilde{\boldsymbol{s}}=\boldsymbol{s}$. This gives rise to the following contextual stochastic optimization problem:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{s}}=\boldsymbol{s}\right]. \tag{1}$$

The core idea of contextual optimization is to leverage the observed side information $\boldsymbol{s}$ to inform the decision $\boldsymbol{x}$, recognizing the fact that the distribution of the uncertain parameter $\tilde{\boldsymbol{\xi}}$ may depend on it. By minimizing the conditional expectation, the decision-maker obtains a policy that is optimally tailored to the specific context provided by $\boldsymbol{s}$. Unfortunately, the true conditional distribution

$\mathbb{P}(\tilde{\boldsymbol{\xi}}|\tilde{\boldsymbol{s}}=\boldsymbol{s})$ is rarely accessible to decision-makers in practice. Instead, they must rely on historical data $\{(\boldsymbol{s}_n,\boldsymbol{\xi}_n)\}_{n\in[N]}$ to learn this conditional distribution and then optimize for the conditional expectation. Existing approaches in the literature employ various techniques for this learning task. Non-parametric methods, such as Nadaraya-Watson kernel regression (Nadaraya 1964, Watson 1964) or k-nearest neighbors (Fix 1985), offer flexibility without assuming a specific functional form for the conditional distribution. However, these methods can suffer from the "curse of dimensionality," performing poorly when the dimension of $\boldsymbol{s}$ is high (Srivastava et al. 2019, Wang et al. 2024). On the other hand, parametric regression methods usually presume a functional relationship between the uncertain parameter and the contextual information, e.g., $\tilde{\boldsymbol{\xi}}=g(\boldsymbol{s})+\tilde{\boldsymbol{\epsilon}}$. While these methods can yield better convergence results, such strong structural assumptions may not hold in many real-world problems, and may struggle to capture complex multimodal data distributions.

Driven by the limitations of existing learning techniques and inspired by the inherent characteristics of real-world data in various operational contexts, we propose a novel approach that models the joint distribution of the uncertain parameter $\tilde{\boldsymbol{\xi}}$ and the contextual information $\tilde{\boldsymbol{s}}$ using a Gaussian Mixture Model (GMM). This modeling choice offers a compelling balance between the flexibility of non-parametric methods and the structural advantages of parametric models. In many operations management and statistical analysis problems, the Gaussian distribution serves as a cornerstone assumption due to its tractability as well as its empirical adequacy in capturing the characteristics of a wide range of real-world data. The GMM, composed of multiple Gaussian components, can be seen as a natural extension of the Gaussian distribution. At one extreme, the GMM with a single cluster reduces to the classical Gaussian model. At the other extreme, when the number of clusters reaches the number of data points, GMM reduces to the non-parametric kernel density estimation method (Silverman 1986), effectively bridging the gap between parametric and non-parametric approaches.

More importantly, GMMs are particularly adept at capturing complex, multi-modal data distributions and identifying underlying "hidden states" or "regimes" within the data, a feature highly relevant to contextual decision-making. This ability enables the model to adapt its understanding of uncertainty to different contexts based on the observation of the side information. For example, in the financial market, empirical studies have long observed that asset returns exhibit characteristics like skewness and multimodality, suggesting the presence of multiple underlying states or regimes (Beedles 1986, Fabozzi et al. 2005, Schaller and Norden 1997). To address the deviation from the Gaussian distribution, Kon (1984) shows that a handful of Gaussian mixture components can accurately approximate the distribution of stock market returns. More recently, Botte and Bao (2021) from Two Sigma study the integration of side information, such as interest rates and exchange rates, with financial market returns into a GMM framework. Their empirical analysis demonstrates that when contextual information is included, GMMs can effectively identify and model significant financial market regimes throughout history.

A key advantage of adopting the GMM in contextual optimization is that it yields an analytical expression for the conditional distribution. Specifically, if $(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{s}})$ follows a GM distribution $\mathbb{M}$, then the conditional distribution is also a GM distribution with parameters that can be explicitly derived from the parameters of $\mathbb{M}$ (Wang et al. 2022). With this closed-form expression of the conditional distribution, the contextual optimization problem can be reformulated into a structured problem that is amenable to efficient solution. Beyond this analytical tractability, the GMM framework fits nicely into contextual optimization due to its learning property. Contrary to the 'curse of dimensionality' issue often encountered in non-parametric methods, learning GM distributions can sometimes become easier in higher dimensions (Anderson et al. 2014). Intuitively, this is because distributions that are inseparable or overlapping in a low-dimensional space may become separable in a higher-dimensional space. This phenomenon aligns closely with the core idea of contextual optimization: when considering only the uncertain parameter $\tilde{\boldsymbol{\xi}}$, the decision-maker might overlook important information to fully understand its behavior or depict its distribution. However, by introducing the contextual information $\tilde{\boldsymbol{s}}$, the increased dimensionality provides a richer and more accurate characterization of the uncertainty, thereby leading to improved decision quality.

The major contribution of this paper can be summarized as follows:

1. **A GMM framework for contextual optimization**: We propose a GMM framework for contextual optimization where the uncertain parameter and contextual information jointly follow a GM distribution. This framework offers a structured approach to leverage side information by modeling the conditional distribution in an analytical expression. Additionally, we provide a theoretical analysis of the approximation quality for the empirical GMM estimation. Our result demonstrates that the approximation error of the empirical GMM is only linearly dependent on its parameter estimation errors. Under reasonable assumptions, learning GMM parameters to a certain accuracy only requires polynomially many samples. This favorable sample complexity contrasts with the "curse of dimensionality" often inherent in non-parametric methods, where achieving similar approximation quality typically requires a number of samples that grows exponentially with the dimension of the contextual information.
2. **Handling general distributions via normalizing flows**: While our GMM framework provides strong modeling power, real-world data may originate from more complex, or even intractable distributions that cannot be perfectly captured by a pure GMM. To extend the applicability of our framework to such general data distributions, we propose a novel approach employing normalizing flows. Inspired by its representational ability for complex distributions, we train a normalizing flow to transform the original random vector into a latent space, where

the transformed random vector follows a Gaussian or a GM distribution. This allows us to implicitly work with our GMM framework in the latent space, preserving analytical advantages for optimization even when the original data distribution is complex. To the best of our knowledge, this is the first work combining normalizing flows with the analytical tractability of GMMs for approximating general conditional distributions in stochastic optimization.

3. **Distributionally robust optimization**: To mitigate the overfitting issue from empirical GMM estimations and improve decision reliability, we integrate our framework with Distributionally Robust Optimization (DRO). Leveraging the analytical availability of our GMM framework, we construct a Wasserstein ambiguity set centered around the empirical conditional distribution. This formulation avoids the degeneracy issues encountered by many conditional DRO models when they condition on a measure-zero singleton $\{\boldsymbol{s}\}$. We further derive distribution coverage results and establish a performance guarantee for the DRO model. Finally, we extend our DRO framework to handle situations where the true number of components in the underlying GMM is unknown. By considering an expanded ambiguity set that encompasses candidate conditional GM distributions with different plausible cluster numbers, our framework offers performance guarantees against potential misspecification of the model complexity.
4. **Multistage setting**: Building upon the structural insights provided by our GMM framework for conditional distributions, we further develop a novel approximation scheme tailored for multistage stochastic optimization problems with Markovian uncertainty. Existing solution schemes mainly rely on Sample Average Approximation (SAA) or kernel regression to handle the evolving conditional distributions. However, the SAA method involves sequential conditional Monte Carlo sampling whose problem size grows exponentially with the time horizon T, while kernel regression methods also incur a sample complexity that grows exponentially with the dimension of the uncertain parameter. In contrast, our proposed GMM-based approximation scheme offers an attractive alternative whose sample complexity grows only linearly with the time horizon T and does not suffer from the curse of dimensionality issue, rendering the effectiveness of our approach for large-scale high-dimensional multistage problems.

1.1. Literature Review

Contextual optimization: The increasing availability of data correlated with uncertain outcomes has spurred significant research into contextual optimization. As formalized in (1), contextual optimization problems aim to optimize a decision based on the conditional distribution of uncertain parameters given observed side information. As highlighted in the survey by Sadana et al. (2025), the literature on solving data-driven contextual optimization problems can be broadly categorized into three paradigms: decision rule approach, sequential learning and optimization, and integrated learning and optimization. We review the relevant literature within these frameworks.

The decision rule approach directly seeks a policy from the covariate $\boldsymbol{s}$ to the decision $\boldsymbol{x}$ by minimizing an empirical risk or a related objective over the historical data. Various functional forms have been explored for these decision rules. The linear decision rule approach, favored for its interpretability and tractability, was studied by Ban and Rudin (2019) in newsvendor problems. Unfortunately, linear decision rules may not achieve asymptotic optimality for general problems. To obtain greater flexibility, Bertsimas and Koduri (2022) propose to approximate the optimal policy with a linear policy within the reproducing kernel Hilbert space (RKHS). More complex non-linear decision rules have also been implemented, including decision tree (Bertsimas et al. 2019) and neural networks (Zhang and Gao 2017, Huber et al. 2019, Oroojlooyjadid et al. 2020). Finally, recognizing that minimizing empirical risk can be prone to overfitting, particularly with limited data or complex functional classes, distributionally robust variants have been designed to offer robustness against data uncertainty. While the decision rule approach is very computationally efficient at the decision stage, it still faces several challenges. First, the decision rule method typically restricts the potential solution space to the chosen class of decision rules. Additionally, ensuring that the learned decision rule outputs decisions can be difficult. Standard empirical risk minimization does not inherently enforce feasibility, and various problem-specific techniques are needed to project the rule's output onto the feasible region (Zhang et al. 2021, Chen et al. 2023).

The sequential learning and optimization paradigm, also known as predict-then-optimize or prescriptive analytics, follows a two-stage process (Bertsimas and Kallus 2020). In the first stage, a model is trained to predict the conditional distribution or sufficient statistics (e.g., conditional mean and covariance) given the context, while in the second stage, the decision-maker simply solves an optimization problem based on the predicted distribution or statistics. There exist various solution schemes for the prediction stage. Non-parametric methods such as Nadaraya-Watson kernel regression (Nadaraya 1964, Watson 1964) and k-nearest neighbors (Fix 1985) offer flexibility without strong distributional assumptions. However, they can suffer from the curse of dimensionality, performing poorly when the dimension of the contextual information is high (Srivastava et al. 2019, Wang et al. 2024). Parametric regression models, by assuming a specific functional form for the relationship between the contextual information $\boldsymbol{s}$ and $\boldsymbol{\xi}$, achieve a faster convergence rate (Sen and Deng 2018, Kannan et al. 2020a). However, such a functional form assumption may not hold in many practical problems and may struggle to learn complex multimodal data distributions. A key challenge in the sequential learning and optimization framework is the "optimizer's curse," where small errors in the prediction model can lead to significantly suboptimal decisions in the downstream optimization problem. To mitigate these issues, robust optimization, distributionally robust optimization, and variance regularization techniques have been applied to enhance models' performance in the out-of-sample circumstances (Srivastava et al. 2019, Bertsimas and Van Parys 2017, Chenreddy et al. 2022b, Kannan et al. 2020b). Our work belongs to this category and contributes to the literature by its ability to handle complicated high-dimensional multimodal distributions.

The integrated learning and optimization framework follows an end-to-end approach that trains the prediction model based on the downstream task loss (Donti et al. 2017, Elmachtoub and Grigas 2022, Qi et al. 2023). Compared with using a decision-blind loss such as mean squared error (MSE), the end-to-end model can yield a prediction that aligns better with the decision stage. For linear programs, the gradient of the decision vector with respect to the predicted cost vector is either nonexistent or zero. To address this issue, surrogate loss functions and other smoothing techniques are employed to create differentiable proxies for the optimization objective (Wilder et al. 2019, Blondel et al. 2020, Elmachtoub and Grigas 2022, Huang and Gupta 2024). While conceptually appealing for its potential to yield high-quality decisions by directly minimizing over the task loss, end-to-end models can be significantly more computationally demanding and complex to implement compared with the previous two paradigms, particularly for large-scale or combinatorial optimization problems (Tang and Khalil 2024).

We remark that contextual optimization is a rapidly growing area, and it is impossible to review all related studies within this section. Therefore, we refer readers who are interested in this topic to a comprehensive review by Sadana et al. (2025).

Normalizing flows: Normalizing flow is a powerful class of generative models that allow for both efficient sampling and exact likelihood evaluation (Dinh et al. 2014, Rezende and Mohamed 2015, Kobyzev et al. 2020, Papamakarios et al. 2021). A normalizing flow defines a complex probability distribution by applying a sequence of invertible and differentiable transformations (diffeomorphisms) to a simple base distribution (e.g., Gaussian). The probability density of a sample then can be computed using the change of variables formula, which involves the density of the transformed sample under the base distribution and the determinant of the Jacobian of the inverse transformation (Villani et al. 2008). Various normalizing flow architectures have been developed with different properties and objectives. Early examples include simple elementwise flows and linear flows (Dinh et al. 2014). More expressive non-linear transformations were introduced in planar and radial flows (Rezende and Mohamed 2015). Borrowing ideas from residual networks, residual flows offer another way to construct invertible mappings, where the residual connection can be viewed as a discretization of a first-order ordinary differential equation (Chen et al. 2019). Continuous flows move a step forward by directly learning the continuous dynamical system, which is also known as infinitesimal flows (Chen et al. 2018a, Grathwohl et al. 2018).

Among the most successful and widely adopted normalizing flow architectures are coupling flows and autoregressive flows, which maintain a balance between representation power and computational efficiency (Kingma et al. 2016, Papamakarios et al. 2017). Coupling flows partition the input and apply transformations to one part conditioned on the other (Dinh et al. 2014, 2016). Autoregressive flows, which are particularly relevant to our work, structure the transformation such that each output dimension depends only on its previous dimensions in a fixed ordering (Kingma et al. 2016, Papamakarios et al. 2017, Huang et al. 2018). This enables the projection of the contextual information $\boldsymbol{s}$ into the latent space without the realization of uncertain parameters $\boldsymbol{\xi}$. Additionally, this ordering structure results in a triangular Jacobian matrix, making its determinant easily computable. Finally, the universality property of autoregressive flows has been proven, indicating their ability to learn any target probability density (Huang et al. 2018, Jaini et al. 2019). Prominent examples include masked autoregressive flows (Papamakarios et al. 2017), neural autoregressive flows (Huang et al. 2018), and inverse autoregressive flow (Kingma et al. 2016).

Multistage stochastic optimization problem with Markovian uncertainty: Multistage stochastic programming (MSP) problems involve a sequence of decisions made over time in the face of evolving uncertainty (Shapiro et al. 2021). These problems are significantly more complex than the single or two-stage problems due to the need to determine policies that are functions of the revealed uncertain parameters at each stage (Birge and Louveaux 2011). We are particularly interested in a class of MSP problems where the uncertain process follows a Markovian structure, i.e., the distribution of future uncertainty depends only on the current state (Kallenberg 1997). This Markovian property directly fits into the setting of contextual optimization and is prevalent in many real-world applications, such as financial modeling and inventory control.

A standard approach for solving MSP problems is the SAA method (Kleywegt et al. 2002). It assumes the knowledge of the true distribution, and employs a discretization scheme to approximate the MSP problem by a scenario tree generated by conditional Monte Carlo Sampling. However, the size of the scenario tree grows exponentially as the planning horizon increases, making the problem computationally intensive (Shapiro 2003, Reaiche 2016, Jiang and Li 2021, Shapiro et al. 2021). To alleviate this issue, kernel regression methods have also been designed to solve MSP problems with Markovian uncertainty (Park et al. 2022). Although the problem size of kernel regression methods does not grow exponentially with T, it suffers from the curse of dimensionality with respect to the dimension of the uncertain parameter. Hence, there is merit in designing a new solution scheme that offers a better sample complexity.

1.2. Paper Structure and Notation

The remainder of the paper is organized as follows. In Section 2, we formally introduce our Gaussian Mixture Model framework for contextual optimization, providing a theoretical analysis of the approximation quality. We further present a novel approach leveraging normalizing flows to extend the applicability of our framework to general data distributions that may not strictly follow a GMM. Section 3 proposes to address parameter estimation uncertainty and potential misspecification of the cluster number using distributionally robust optimization. Section 4 extends our approach to MSP problems under Markovian uncertainty, establishing a novel GMM-based approximation scheme. We present extensive numerical experiments on synthetic and real-world datasets to demonstrate the practical effectiveness of our proposed approach in Section 5.

Notation We use boldface letters to denote vectors and matrices. Random variables are indicated with a tilde (e.g., $\tilde{\boldsymbol{\xi}}$), while their realizations appear without the tilde (e.g., $\boldsymbol{\xi}$). The multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and its density function is written as $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We denote the vector of all ones by $\mathbf{e}$. The probability simplex in $\mathbb{R}^K_+$ is denoted as Δ^K. The little-o notation $o(\cdot)$ is used for asymptotic analysis: $h(\epsilon) = o(\epsilon)$ if $h(\epsilon)/\epsilon \to 0$, that is, $h(\epsilon)$ becomes negligible relative to ϵ as $\epsilon \to 0$.

2. Contextual Stochastic Optimization with Mixture Models

We consider the contextual stochastic optimization problem given by

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{M}} \left[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \middle| \tilde{\boldsymbol{s}} = \boldsymbol{s} \right]. \tag{2}$$

The random vector $(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{\xi}}) \in \mathbb{R}^{Q+R}$ follows a joint distribution

$$\mathbb{M} := \sum_{k\in[K]} p^k \mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k),$$

which is a Gaussian mixture (GM) distribution with component means and covariances given by

$$\boldsymbol{\mu}^k := \begin{bmatrix} \boldsymbol{\mu}^k_{\boldsymbol{s}} \\ \boldsymbol{\mu}^k_{\boldsymbol{\xi}} \end{bmatrix} \in \mathbb{R}^{Q+R} \quad \text{and} \quad \boldsymbol{\Sigma}^k := \begin{bmatrix} \boldsymbol{\Sigma}^k_{\boldsymbol{ss}} & \boldsymbol{\Sigma}^k_{\boldsymbol{s\xi}} \\ \boldsymbol{\Sigma}^k_{\boldsymbol{\xi s}} & \boldsymbol{\Sigma}^k_{\boldsymbol{\xi\xi}} \end{bmatrix} \in \mathbb{S}^{Q+R}_+ \qquad \forall k \in [K].$$

The mixture weights are given by $\boldsymbol{p} \in \Delta^K$, where $p^k \in [0,1]$ represents the weight of the k-th component. At first glance, problem (2) may appear challenging because it involves computing the conditional expectation of $\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ given that $\tilde{\boldsymbol{s}} = \boldsymbol{s}$ is observed. However, a recent result establishes that the conditional distribution of a GM distribution remains a GM distribution, thereby enabling the development of a principled solution scheme.

Lemma 1 **(Lemma 1 in Wang et al. (2022))**. *The conditional distribution of $\tilde{\boldsymbol{\xi}}$ given $\tilde{\boldsymbol{s}} = \boldsymbol{s}$ is a GM*

$$\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}} := \sum_{k\in[K]} p^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathbb{N}(\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}),$$

where

$$\begin{aligned} \boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}} &= \boldsymbol{\mu}^k_{\boldsymbol{\xi}} + \boldsymbol{\Sigma}^k_{\boldsymbol{\xi s}} (\boldsymbol{\Sigma}^k_{\boldsymbol{ss}})^{-1} (\boldsymbol{s} - \boldsymbol{\mu}^k_{\boldsymbol{s}}), \\ \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}} &= \boldsymbol{\Sigma}^k_{\boldsymbol{\xi\xi}} - \boldsymbol{\Sigma}^k_{\boldsymbol{\xi s}} (\boldsymbol{\Sigma}^k_{\boldsymbol{ss}})^{-1} \boldsymbol{\Sigma}^k_{\boldsymbol{s\xi}}, \text{ and} \\ p^k_{\boldsymbol{\xi}|\boldsymbol{s}} &= \frac{p^k \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)}{\sum_{j=1}^K p^j \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^j_{s}, \boldsymbol{\Sigma}^j_{\boldsymbol{ss}}\right)}. \end{aligned} \tag{3}$$

This result simplifies problem (2) into an *un*conditional stochastic optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]. \tag{4}$$

For specific loss functions, such as the mean-variance utility used in portfolio optimization, problem (4) can be solved directly by plugging the conditional parameters. For more general loss functions, problem (4) is amenable to a high-quality solution via the sample average approximation. In this approach, the expectation is replaced with the average of the loss function over samples drawn from the conditional distribution $\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}$.

In practice, however, the joint distribution $\mathbb{M}$ is rarely available to the decision maker. Instead, one must rely on N historical observations $\{(\boldsymbol{s}_n,\boldsymbol{\xi}_n)\}_{n\in[N]}$ to estimate these parameters. Under this data-driven setting, a standard approach is to fit a GM distribution $\hat{\mathbb{M}}$ to these observations, and solve the approximate contextual stochastic optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right],$$

where the approximate conditional distribution $\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}$ is obtained by applying Lemma 1 to $\hat{\mathbb{M}}$. While this empirical approach is straightforward to implement, the significance of its approximation error is still not fully understood. In the following section, we provide a theoretical analysis of this empirical approach to quantify its approximation quality.

2.1. Approximation Quality

We analyze the approximation quality of this empirical approximation. Quantifying this approximation error is crucial for understanding the reliability of the empirical approach. To facilitate our analysis, we make the following standard assumptions:

(A) The loss function is bounded, i.e., $-\overline{\ell}\le\ell(\boldsymbol{x},\boldsymbol{\xi})\le\overline{\ell}$ for some $\overline{\ell}\in\mathbb{R}_+$.

(B) There exists a constant $\gamma\in\mathbb{R}_+$ such that $\|\boldsymbol{\mu}^k\|,\|\hat{\boldsymbol{\mu}}^k\|\le\gamma$. In addition, the covariance matrices $\boldsymbol{\Sigma}^k$ and $\hat{\boldsymbol{\Sigma}}^k$ are positive definite, and there exist constants $\alpha,\beta\in\mathbb{R}_{++}$ such that $\alpha\mathbb{I}\preceq\boldsymbol{\Sigma}^k,\hat{\boldsymbol{\Sigma}}^k\preceq\beta\mathbb{I}$ for all $k\in[K]$.

(C) There exists a constant $\underline{p}\in\mathbb{R}_{++}$ such that $p^k\ge\underline{p}$ for all $k\in[K]$.

The requirement that the loss function is bounded in Assumption (A) is made to simplify the exposition. It can be relaxed to the assumption that the loss function is Lebesgue integrable with respect to a Gaussian measure, which is satisfied whenever the function has polynomial growth. The requirement $\boldsymbol{\Sigma}^k\succeq\alpha\mathbb{I}\succ\mathbf{0}$ in Assumption (B) is without loss of generality because we can always consider the affine subspace where the Gaussian distribution is supported. The constant $\underline{p}$ in Assumption (C) is known as the condition number of the GMM (Kalai et al. 2012). This assumption is standard in the GMM literature and is typically imposed to ensure the feasibility of learning from data.

Theorem 1. *Let $\mathbb{M} = \sum_{k\in[K]} p^k \mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$ be the true underlying Gaussian mixture distribution and $\hat{\mathbb{M}} = \sum_{k\in[K]} \hat{p}^k \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)$ be the estimated one. If $|p^k - \hat{p}^k| \le \epsilon_p$, $\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k| \le \epsilon_{\boldsymbol{\mu}}$, and $\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \le \epsilon_{\boldsymbol{\Sigma}}$ for all $k \in [K]$, then*

$$\left|\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] - \mathbb{E}_{\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]\right| \le \overline{\ell}\left(2\left(C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right) + c_R\|\boldsymbol{s}\| + c'_R\right), \tag{5}$$

where $C_Q = \left(1 + \frac{\epsilon_p}{\underline{p}}\right)\frac{Q\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}$, $C'_Q = \left(1 + \frac{\epsilon_p}{\underline{p}}\right)\left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha} + \frac{Q\epsilon_{\boldsymbol{\Sigma}}\gamma}{\alpha^2}\right)$, *and* $C''_Q = \frac{\epsilon_p}{\underline{p}} + \left(1 + \frac{\epsilon_p}{\underline{p}}\right)\left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha}\gamma + \frac{Q\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}(\gamma^2 + \alpha) + o(\epsilon_\mu + \epsilon_{\boldsymbol{\Sigma}})\right)$, $c_R = \frac{\sqrt{R}(\alpha+\beta)}{\alpha^3}\epsilon_{\boldsymbol{\Sigma}}$, *and* $c'_R = \frac{\sqrt{R}}{\alpha}\left(\left(\frac{\beta}{\alpha} + 1\right)\epsilon_{\boldsymbol{\mu}} + \frac{\alpha+\beta}{\alpha^2}\gamma\epsilon_{\boldsymbol{\Sigma}}\right) + \frac{1}{2}\frac{R^2}{\alpha}\left(\frac{\beta}{\alpha}\right)^2\epsilon_{\boldsymbol{\Sigma}}$.

The bound accounts for the errors arising from the estimation of mixture probabilities, given by $2\overline{\ell}\left(C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right)$, and from the estimation of the mean and covariance of each Gaussian component, captured by $\overline{\ell}(c_R\|\boldsymbol{s}\| + c'_R)$. We observe that the approximation quality deteriorates as α decreases and γ increases. These constants are introduced to conservatively upper bound $\|(\boldsymbol{\Sigma}^k_{\boldsymbol{ss}})^{-1}\| \le \frac{1}{\alpha}$ and $\|\boldsymbol{s} - \boldsymbol{\mu}^k_s\| \le \|\boldsymbol{s}\| + \gamma$. Thus, when there is no Gaussian component in which the covariates $\tilde{\boldsymbol{s}}$ exhibit high variability or whose mean is close to $\boldsymbol{s}$, the approximation error becomes significant. This corresponds to cases where samples $\tilde{\boldsymbol{\xi}}$ in the vicinity of $\tilde{\boldsymbol{s}} = \boldsymbol{s}$ are rarely observed, which slows the learning rate of the conditional distribution. The dependence in $\|\boldsymbol{s}\|$ stems from the same reasoning: since the GM distribution is sub-Gaussian, the density $f(\boldsymbol{s})$ decreases as the covariates $\boldsymbol{s}$ move further from the origin. The bound also depends on β, with the proof showing that the error increases as $\|\boldsymbol{\Sigma}^k_{\boldsymbol{\xi s}}\| \le \beta$ grows. This reflects strong dependence between $\tilde{\boldsymbol{s}}$ and $\tilde{\boldsymbol{\xi}}$, which complicates the estimation of the conditional distribution. In contrast, in the limiting case of zero correlation, the marginal distribution of $\tilde{\boldsymbol{\xi}}$ alone suffices, making learning easier.

Finally, this result confirms that the error introduced by the data-driven approximation is linear in ϵ_p, $\epsilon_{\boldsymbol{\mu}}$, and $\epsilon_{\boldsymbol{\Sigma}}$. This linear dependency is particularly significant when considering the sample complexity of GMM learning. Under standard regularity conditions, GMM algorithms typically yield parameter estimation errors that decay polynomially in the sample size N and the dimension $Q + R$ (See Remark 1 for details). This stands in stark contrast to existing data-driven methods for general contextual stochastic optimization, which often suffer from approximation errors that decay slowly with the dimension Q of the contextual covariates $\tilde{\boldsymbol{s}}$ (Srivastava et al. 2021, Wang et al. 2024). Moreover, in the special case of a single Gaussian component ($K = 1$), standard concentration results imply that $\epsilon_{\boldsymbol{\mu}}$ and $\epsilon_{\boldsymbol{\Sigma}}$ are of the order $O(\frac{1}{\sqrt{N}})$.

Remark 1. Learning mixtures of Gaussians is a highly active research area, where numerous algorithms have been developed, accompanied by theoretical results that establish their sample complexity. The work (Moitra and Valiant 2010, Kalai et al. 2012) develops an algorithm that, for any GM distribution $\mathbb{M} = \sum_{k\in[K]} p^k \mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$, outputs an ϵ-estimate $\hat{\mathbb{M}} = \sum_{k\in[K]} p^k \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)$

satisfying $|p_k - \hat{p}_k| \leq \epsilon$ and $\mathbb{TV}(\mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)) \leq \epsilon$ with high probability. Here, the total variation distance is used to determine the closeness of the component Gaussians, which is defined as $\mathbb{TV}(f, \hat{f}) = \int (f(\boldsymbol{\xi}) - \hat{f}(\boldsymbol{\xi}))\mathrm{d}\boldsymbol{\xi}$ for any two distributions f and $\hat{f}$. In the case of two Gaussians $\mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$ and $\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k))$, a small total variation between them implies that the parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ have small estimation errors as well (see Proposition 4). The running time and data requirement of this algorithm are polynomial in the dimension D, accuracy ϵ, and condition number $\kappa = \min\{\underline{p}, \min_{i,j\in[K]:i\neq j} \mathbb{TV}(\mathbb{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i), \mathbb{N}(\boldsymbol{\mu}^j, \boldsymbol{\Sigma}^j))\}^{-1}$. Hence, the number of samples N required to sustain an approximation error of at most τ in (5) will be polynomial in the desired accuracy τ and problem dimensions Q and R. More recent literature provides tighter bounds for various problem settings and algorithms. For instance, when the number of components $K = 2$, Hardt and Price (2015) proposes an algorithm that can learn an arbitrary GM distribution to a certain accuracy with polynomially many samples. For a spherical GM distribution that is $\Omega(\sqrt{\log k})$ separable, Kwon and Caramanis (2020) shows that $\tilde{O}(K(Q+R)/\epsilon^2)$ samples suffice for the Expectation Maximization (EM) algorithm to achieve ϵ parameter estimation accuracy. Using the PCA learning method, Ashtiani et al. (2018b) derive a sample-efficient learning algorithm that can learn an arbitrary GM distribution to certain TV accuracy with $\tilde{O}(K(Q+R)^2/\epsilon^4)$ samples, and an axis-aligned GM distribution with $\tilde{O}(K(Q+R)/\epsilon^4)$ samples. This result is later improved by their subsequent work (Ashtiani et al. 2018a) with a sample complexity of $\tilde{O}(K(Q+R)^2/\epsilon^2)$ for general GM distributions, and $\tilde{O}(K(Q+R)/\epsilon^2)$ for axis-aligned GMs. For high dimensional GM distributions, Kwon and Caramanis (2020) show that with a suitable initialization and $N \geq \tilde{O}((\min_{k\in[K]} p^{k\star})^{-1}(Q+R)/\epsilon^2)$, the Expectation Maximization (EM) algorithm converges in $T = O(\log(1/\epsilon))$ iterations to estimates $\hat{p}^k$, $\hat{\mu}^k$, $(\hat{\sigma}^k)^2$ with accuracies $\epsilon_p = \max_{k\in[K]} p^{k\star}\epsilon$, $\epsilon_\mu = \max_{k\in[K]} \sigma^{k\star}\epsilon$, $\epsilon_\Sigma = (\max_{k\in[K]} \sigma^{k\star})^2\epsilon/\sqrt{Q+R}$, respectively.

So far, we have introduced our GMM framework for contextual optimization and provided a theoretical analysis of the approximation quality. However, real-world data may originate from more complex distributions that are not strictly GMMs. In the next section, we extend the applicability of our framework to general distributions.

2.2. Handling General Probability Distributions

While GMMs possess expressive power that can model many real-world distributions fairly effectively, it may not accurately capture the precise characteristics of these distributions. Our solution scheme can, in principle, be extended to a mixture of parametric distributions since their conditional distributions are also closed-form, which can be tractable (Cambanis et al. 1981). However, in some situations, these assumptions remain restrictive, and the data may be more faithfully represented as (a mixture of) non-elliptical distributions. In this work, we strive to handle even more general distributions while still exploiting the unique property of GMM in computing the conditional distributions in closed form. We propose employing normalizing flows (Dinh et al. 2014, Rezende and Mohamed 2015), a powerful mechanism for representing complex distributions through a sequence of mappings from a simple base distribution, such as a Gaussian or a mixture of Gaussians. To our knowledge, the use of normalizing flows and Gaussian mixture models to compute conditional distributions of an intractable distribution has not previously been considered in the literature.

Let $(\tilde{\boldsymbol{s}}', \tilde{\boldsymbol{\xi}}') \sim \mathbb{P}$ be the data-generating random vector and $(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{\xi}}) \sim \mathbb{M}$ be the latent random vector governed by a tractable base distribution $\mathbb{M}$, which in our case is set to a mixture of Gaussians. The framework seeks a differentiable and invertible function $T \in \mathbb{R}^{Q+R} \to \mathbb{R}^{Q+R}$ such that $T(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{\xi}})$ follows the distribution $\mathbb{P}$, which enables density estimation and sampling via the base distribution $\mathbb{M}$. Under reasonable assumptions, normalizing flows are extremely expressive: for any distribution $\mathbb{P}$, there always exists a suitable transformation T that brings $\mathbb{M}$ into $\mathbb{P}$, even if the base distribution is described by a simple uniform distribution $\mathbb{M} = \mathbb{U}[0,1]^{Q+R}$. In practice, the function is typically restricted to a neural network representation $T_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$, enabling large-scale training with state-of-the-art algorithms. The normalizing flow framework has been successfully implemented in various large-scale tasks, including image generation (Ho et al. 2019), video generation (Kumar et al. 2019), reinforcement learning (Mazoure et al. 2020, Touati et al. 2020), and computer graphics (Müller et al. 2019). We refer the reader to the survey papers (Kobyzev et al. 2020, Papamakarios et al. 2021) for comprehensive reviews of normalizing flows.

Since T is a *diffeomorphism*, i.e., differentiable and invertible, the pushforward density function can be evaluated in closed form via the change-of-variables formula (Rudin 1987, Theorem 7.24):

$$f_{\mathbb{P}}(\boldsymbol{s}', \boldsymbol{\xi}') = f_{\mathbb{M}}\left(T^{-1}(\boldsymbol{s}', \boldsymbol{\xi}')\right) |\det \mathbf{J}_{T^{-1}}(\boldsymbol{s}', \boldsymbol{\xi}')|,$$

where $\det \mathbf{J}_{T^{-1}}(\boldsymbol{s}', \boldsymbol{\xi}')$ denotes the Jacobian determinant of T^{-1} at $(\boldsymbol{s}', \boldsymbol{\xi}')$. To find the best neural network parameters $\boldsymbol{\theta}$, we minimize the KL-divergence between the target density $f_{\mathbb{P}}$ and the pushforward of the base density $f_{\mathbb{M}}$ through the inverse transformation $T_{\boldsymbol{\theta}}^{-1}$:

$$\min_{\boldsymbol{\theta}} \mathscr{D}\left(f_{\mathbb{P}}(\boldsymbol{s}', \boldsymbol{\xi}') \,||\, f_{\mathbb{M}}\left(T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}', \boldsymbol{\xi}')\right) |\det \mathbf{J}_{T_{\boldsymbol{\theta}}^{-1}}(\boldsymbol{s}', \boldsymbol{\xi}')|\right).$$

By the definition of the KL divergence, this is equivalent to the following optimization problem:

$$\min_{\boldsymbol{\theta}} -\mathbb{E}_{\mathbb{P}}\left[\log f_{\mathbb{M}}\left(T_{\boldsymbol{\theta}}^{-1}(\tilde{\boldsymbol{s}}', \tilde{\boldsymbol{\xi}}')\right)) + \log |\det \mathbf{J}_{T_{\boldsymbol{\theta}}^{-1}}(\tilde{\boldsymbol{s}}', \tilde{\boldsymbol{\xi}}')|\right].$$

Using the observed samples $\{(\boldsymbol{s}_n, \boldsymbol{\xi}_n)\}_{n\in[N]}$, the expectation in the objective is then approximated using the sample-average approximation

$$\min_{\theta} -\frac{1}{N}\sum_{n\in[N]}\left(\log f_{\mathbb{M}}\left(T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}'_n, \boldsymbol{\xi}'_n)\right)) + \log|\det \mathbf{J}_{T_{\boldsymbol{\theta}}^{-1}}(\boldsymbol{s}'_n, \boldsymbol{\xi}'_n)|\right),$$

which is solved at scale using stochastic gradient descent.

In this work, we further assume a separable structure in the transformation

$$(\boldsymbol{s}', \boldsymbol{\xi}') = T_{\boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{\xi}) = (T_{\boldsymbol{\theta},\boldsymbol{s}'}(\boldsymbol{s}), T_{\boldsymbol{\theta},\boldsymbol{\xi}'}(\boldsymbol{s}, \boldsymbol{\xi})),$$

which enforces the transformation for contextual covariates $\boldsymbol{s}'$ to depend solely on $\boldsymbol{s}$. This structure is relatively general and includes, as a special case, the popular *autoregressive flows* (Kingma et al. 2016, Papamakarios et al. 2017, Huang et al. 2018). Note that, by construction, the inverse transformation is given by

$$T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}', \boldsymbol{\xi}') = \left(T_{\boldsymbol{\theta},\boldsymbol{s}'}^{-1}(\boldsymbol{s}'), T_{\boldsymbol{\theta},\boldsymbol{\xi}'}^{-1}\left(T_{\boldsymbol{\theta},\boldsymbol{s}'}^{-1}(\boldsymbol{s}'), \boldsymbol{\xi}'\right)\right),$$

where $T_{\boldsymbol{\theta},\boldsymbol{s}'}^{-1}(\boldsymbol{s}')$ is the inverse mapping with respect to $\boldsymbol{s}$, acting as a retransformation for the contextual information, while $T_{\boldsymbol{\theta},\boldsymbol{\xi}'}^{-1}\left(T_{\boldsymbol{\theta},\boldsymbol{s}'}^{-1}(\boldsymbol{s}'), \boldsymbol{\xi}'\right)$ denotes the inverse transformation with respect to $\boldsymbol{\xi}$.

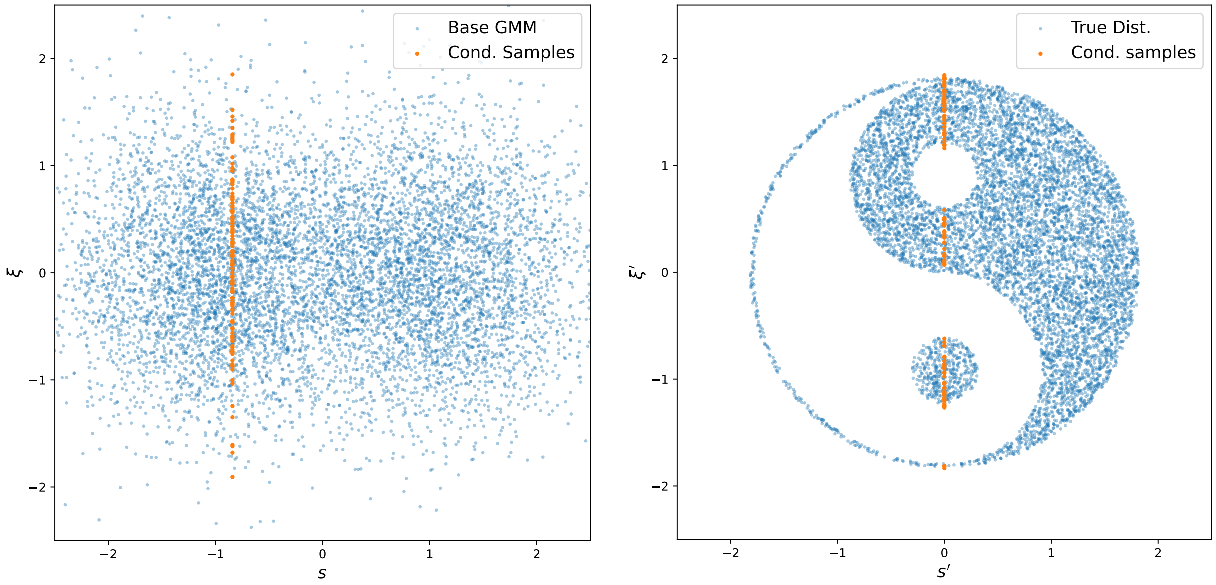
This normalizing flow architecture, including pushforward mapping and inverse mapping, allows for two key benefits. First, the transformations of $\boldsymbol{s}'$ or $\boldsymbol{s}$ are independent of the uncertain parameter, meaning we can transform the contextual information between the original and latent spaces without knowing the realization of $\tilde{\boldsymbol{\xi}}'$. This aligns perfectly with the setup of contextual optimization, where decision-makers can only observe the realization of contextual information when making the decision. Second, the inverse transformation of $\boldsymbol{\xi}$ utilizes the projected contextual information $\boldsymbol{s}$, facilitating the computation of the conditional distribution of $\tilde{\boldsymbol{\xi}}'$ given $\tilde{\boldsymbol{s}}' = \boldsymbol{s}'$ in closed form, as described below.

Proposition 1. *Let $\boldsymbol{s} = T_{\boldsymbol{\theta},\boldsymbol{s}'}^{-1}(\boldsymbol{s}')$. Then the conditional distribution $\mathbb{P}_{\boldsymbol{\xi}'|\boldsymbol{s}'}$ has a density given by*

$$f_{\mathbb{P}}(\boldsymbol{\xi}'|\boldsymbol{s}') = f_{\mathbb{M}}\left(T_{\boldsymbol{\theta},\boldsymbol{\xi}'}^{-1}(\boldsymbol{s}, \boldsymbol{\xi}')\middle|\boldsymbol{s}\right)\left|\det \mathbf{J}_{T_{\boldsymbol{\theta},\boldsymbol{\xi}'}^{-1}}(\boldsymbol{s}, \boldsymbol{\xi}')\right|,$$

where the Jacobian $J_{T_{\boldsymbol{\theta},\boldsymbol{\xi}'}^{-1}}(\boldsymbol{s}, \boldsymbol{\xi}')$ is taken only with respect to $\boldsymbol{\xi}'$.

Proposition 1 illustrates how to obtain the conditional probability density function for a general distribution from a base distribution and a normalizing flow. This result is useful for multistage stochastic problems, which will be illustrated in Section 4. For the single-stage problem, however, the decision maker does not need to explicitly compute the density function. Instead, one can solve the stochastic optimization problem by efficient sampling from the base distribution. We now provide a concrete example below.



(a) Latent Space **(b)** Original Space

Figure 1 A visualization of the conditional sampling generating scheme using normalizing flow.

Example 1. This example illustrates the conditional sampling procedure enabled by our normalizing flow framework. As visualized by Figure 1, the original samples, denoted by blue points in Figure 1(b), come from a complex 'Yin-Yang' shaped distribution[1]. A normalizing flow is then trained to learn an invertible mapping between this true distribution and a base GM distribution in the latent space. Figure 1(a) shows the mapped samples (blue dots) in the latent space.

Suppose now we observe the side information $s' = 0$. Then, using the trained normalizing flow, we can recover the latent covariate representation via $s = T^{-1}_{\boldsymbol{\theta},s'}(s') = -0.84$. Within the latent space, we leverage the GMM's analytical properties. Using the result from Lemma 1, we compute the conditional base distribution in closed form. Since this conditional distribution is also a GMM, an arbitrary number of samples can be efficiently generated, which is denoted by the orange points in Figure 1(a). In the final step, these latent samples are propagated through the forward transformation, yielding the desired conditional samples in the original space.

3. Distributionally Robust Optimization Framework

The GMM framework presented in Section 2, while powerful, relies on empirical estimation from limited historical observations. This can lead to solutions that overfit the training sample, yielding poor performance in out-of-sample scenarios. Furthermore, when learning the GM distribution, the model's structure (the number of mixture components K) may be misspecified. To systematically

[1] Yin and Yang is a Chinese philosophy that views the universe as governed by opposing yet complementary forces observable in nature, like the moon and sun, darkness and light, cold and heat.

address these issues and enhance decision reliability, we integrate our approach with Distributionally Robust Optimization (DRO). In what follows, we will formulate this robust model, establish its reformulation, and derive performance guarantees that also account for potential misspecification of the model complexity.

3.1. Reformulation and Performance Guarantee

The DRO model seeks a solution that hedges against the worst-case distribution within a prescribed ambiguity set, i.e.,

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\varepsilon} \mathbb{E}_{\mathbb{Q}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\big|\tilde{\boldsymbol{s}}=\boldsymbol{s}\right], \tag{6}$$

where the ambiguity set $\mathcal{P}_\varepsilon$ is a neighborhood of radius $\varepsilon\in\mathbb{R}_+$, centered at the estimated $\hat{\mathbb{M}}$, defined with respect to a suitable probability metric. However, such a formulation suffers from degeneracy due to the conditioning on the singleton $\{\boldsymbol{s}\}$, which has zero measure. A common workaround is to condition on a set $\mathcal{S}$ of positive measure that contains $\boldsymbol{s}$ (Van Eekelen 2023). Such a scheme, however, may lead to difficulty in defining the structure and the size of $\mathcal{S}$ in practice. This limitation motivates the development of an alternative formulation that provides both finite-sample performance guarantees and asymptotic consistency.

A key advantage of our GMM framework is that the empirical conditional distribution $\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}$ is analytically available given the empirical estimation of $\mathbb{M}$. Leveraging this, we can directly construct a data-driven ambiguity set centered at this empirical conditional distribution, i.e.,

$$\mathcal{P}_\varepsilon := \left\{\mathbb{Q}\in\mathcal{P}_0 : \mathbb{W}_2\left(\mathbb{Q},\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}\right)\le\varepsilon\right\}. \tag{7}$$

Here, we adopt the type-2 Wasserstein distance $\mathbb{W}_2$ (Kantorovich and Rubinshtein 1958, Gao and Kleywegt 2023) defined as

$$\mathbb{W}_2(\mathbb{Q}_1,\mathbb{Q}_2) := \inf_{\pi\in\Pi(\mathbb{Q}_1,\mathbb{Q}_2)}\left(\int_{\mathcal{Z}\times\mathcal{Z}}\|\boldsymbol{z}_1-\boldsymbol{z}_2\|^2\,\pi(\mathrm{d}\boldsymbol{z}_1,\mathrm{d}\boldsymbol{z}_2)\right)^{\frac{1}{2}},$$

where $\Pi(\mathbb{Q}_1\times\mathbb{Q}_2)$ is the set of all joint probability distributions of random vectors $\tilde{\boldsymbol{z}}_1$ and $\tilde{\boldsymbol{z}}_2$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively. The use of the type-2 Wasserstein distance facilitates theoretical performance guarantees, as we will discuss in the following section. Moreover, compared to the type-1 Wasserstein distance, it avoids pathological worst-case distributions and often yields more robust and higher-quality decisions (Byeon 2025).

Using this ambiguity set, we obtain a distributionally robust optimization problem that minimizes the worst-case *unconditional* expectation with respect to all distributions within a Wasserstein ball centered at the estimated conditional distribution:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\varepsilon} \mathbb{E}_{\mathbb{Q}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]. \tag{8}$$

The practical utility and theoretical performance guarantee of model (8) require coverage of the true distribution. To ensure that $\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}} \in \mathcal{P}_\varepsilon$, the radius ε must be chosen appropriately. The following Theorem demonstrates that the radius required to ensure coverage scales polynomially with the estimation errors ϵ_p, $\epsilon_{\boldsymbol{\mu}}$, and $\epsilon_{\boldsymbol{\Sigma}}$.

THEOREM 2. *Assume the conditions in Theorem 1 hold. Then it is sufficient to set*

$$\varepsilon^2 = \left(\left(\tfrac{\beta}{\alpha}+1\right)\epsilon_{\boldsymbol{\mu}} + \tfrac{\alpha+\beta}{\alpha^2}\left(\|\boldsymbol{s}\|+\gamma\right)\epsilon_{\boldsymbol{\Sigma}}\right)^2 + D\left(\tfrac{\beta}{\alpha}\right)^2\epsilon_{\boldsymbol{\Sigma}} + 2\left(4\gamma^2+2D\beta\right)\left(C_Q\|\boldsymbol{s}\|^2 + C_Q'\|\boldsymbol{s}\| + C_Q''\right). \tag{9}$$

to guarantee that $\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}} \in \mathcal{P}_\varepsilon$.

This result yields the following out-of-sample performance guarantee for the solution to problem (8), respectively.

COROLLARY 1. *Let $\hat{\boldsymbol{x}}$ and $\hat{J}$ be an optimal solution and the optimal value of the distributionally robust optimization problem* (8). *If the radius ε of the ambiguity set $\mathcal{P}_\varepsilon$ is chosen according to* (9), *then the following guarantee holds:*

$$\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] \le \hat{J}.$$

Having established the performance guarantees of our DRO framework, we now turn to its practical implementation. By (Blanchet and Murthy 2019, Remark 1), if the loss function $\ell(\boldsymbol{x},\boldsymbol{\xi})$ is upper semicontinuous in $\boldsymbol{\xi}$, this problem is equivalent to the following minimization problem involving the estimated conditional distribution $\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}$:

$$\min_{\boldsymbol{x}\in\mathcal{X},\lambda\in\mathbb{R}_+} \varepsilon^2\lambda + \mathbb{E}_{\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\sup_{\boldsymbol{\omega}\in\mathbb{R}^R} \ell(\boldsymbol{x},\boldsymbol{\omega}) - \lambda\|\boldsymbol{\omega}-\tilde{\boldsymbol{\xi}}\|^2\right].$$

This formulation is amenable to the sample average approximation:

$$\min_{\boldsymbol{x}\in\mathcal{X},\lambda\in\mathbb{R}_+} \varepsilon^2\lambda + \frac{1}{M}\sum_{m\in[M]} \sup_{\boldsymbol{\omega}\in\mathbb{R}^R} \ell(\boldsymbol{x},\boldsymbol{\omega}) - \lambda\|\boldsymbol{\omega}-\boldsymbol{\xi}_m\|^2, \tag{10}$$

where $\{\boldsymbol{\xi}_m\}_{m\in[M]}$ are samples drawn from the conditional distribution $\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}$. When the loss function is piecewise affine, the resulting DRO problem admits a tractable second-order conic programming (SOCP) reformulation that can be efficiently solved using standard optimization solvers. Although the derivation is straightforward, to our knowledge, it does not appear explicitly in the existing literature. We therefore provide the following result.

PROPOSITION 2. *If the loss function takes the form $\ell(\boldsymbol{x},\boldsymbol{\xi}) = \max_{j\in[J]} \boldsymbol{a}_j(\boldsymbol{x})^\top\boldsymbol{\xi} + b_j(\boldsymbol{x})$, then the distributionally robust optimization problem* (10) *is equivalent to the second-order conic program*

$$\begin{array}{ll}
\min & \varepsilon^2\lambda + \dfrac{1}{M}\displaystyle\sum_{m\in[M]} \gamma_m \\
\text{s.t.} & \boldsymbol{x}\in\mathcal{X},\ \lambda\in\mathbb{R}_+,\ \boldsymbol{\gamma}\in\mathbb{R}^M \\
& \left\|\begin{bmatrix} 2\lambda\boldsymbol{\xi}_m + \boldsymbol{a}_j(\boldsymbol{x}) \\ \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m - b_j(\boldsymbol{x}) - \lambda \end{bmatrix}\right\| \le \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m - b_j(\boldsymbol{x}) + \lambda \quad \forall j\in[J] \quad \forall m\in[M] \\
& \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m \ge b_j(\boldsymbol{x}) \quad \forall j\in[J] \quad \forall m\in[M].
\end{array}$$

Remark 2. The DRO model and its tractable reformulation presented in Proposition 2 can be readily extended to cases where the sample $(\tilde{\boldsymbol{s}}', \tilde{\boldsymbol{\xi}}')$ follows a general, non-GM distribution $\mathbb{P}$. Specifically, the sample average approximation in (10) would be constructed using samples $\{\boldsymbol{\xi}'_m\}_{m\in[M]}$ drawn from the target conditional distribution $\mathbb{P}_{\boldsymbol{\xi}'|\boldsymbol{s}'}$. As detailed in Section 2.2, these samples can be generated efficiently by first drawing latent $\{\boldsymbol{\xi}_m\}_{m\in[M]}$ samples from the conditional base distribution, and then applying the learned forward transformation $\xi' = T_{\boldsymbol{\theta},\boldsymbol{\xi}'}(\boldsymbol{s}, \boldsymbol{\xi})$.

Together, Proposition 2 and Remark 1 demonstrate that our DRO framework is both computationally tractable and broadly applicable to general distributions. In the next part of this section, we will address the challenge of model misspecification.

3.2. Robustifying Against Unknown Mixture Size K

A practical challenge in applying the GMM framework is that the true number of mixture components K is typically unknown and must be chosen as a hyperparameter. Selecting an incorrect number K' can lead to model misspecification, which invalidates the performance guarantees derived in the previous section. Our DRO framework can be naturally extended to hedge against this structural uncertainty. Specifically, although the exact number of K is unknown, the decision maker could generally determine a finite set $\mathcal{K} \subseteq \mathbb{Z}_+$ that constitutes all plausible candidate numbers. In this case, one could train GMM algorithms under different mixture sizes, resulting in $|\mathcal{K}|$ candidate conditional distributions: $\hat{\mathbb{M}}^L_{\boldsymbol{\xi}|\boldsymbol{s}}$ for $L \in \mathcal{K}$. Then, the decision maker can solve the distributionally robust optimization problem (8) using an ambiguity set of an appropriate radius centered at one of these distributions. The following proposition provides a recommendation for designing such an ambiguity set.

Proposition 3. *Define $\overline{\mathbb{W}}_2^2(\hat{\mathbb{M}}^K_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}})$ for $K, K' \in \mathcal{K}$ to be the optimal value of the following linear program:*

$$\begin{aligned}
\overline{\mathbb{W}}_2^2(\hat{\mathbb{M}}^K_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}) := \min\ & \sum_{i\in[K]} \sum_{j\in[K']} \pi_{ij} \left(\|\hat{\boldsymbol{\mu}}^i_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{\boldsymbol{\mu}}^j_{\boldsymbol{\xi}|\boldsymbol{s}}\|^2 + \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}^i_{\boldsymbol{\xi}|\boldsymbol{s}} + \hat{\boldsymbol{\Sigma}}^j_{\boldsymbol{\xi}|\boldsymbol{s}} - 2(((\hat{\boldsymbol{\Sigma}}^i_{\boldsymbol{\xi}|\boldsymbol{s}})^{\frac{1}{2}} \hat{\boldsymbol{\Sigma}}^j_{\boldsymbol{\xi}|\boldsymbol{s}} (\hat{\boldsymbol{\Sigma}}^i_{\boldsymbol{\xi}|\boldsymbol{s}})^{\frac{1}{2}})^{\frac{1}{2}} \right) \right) \\
\text{s.t.}\ & \boldsymbol{\pi} \in \Delta^K \times \Delta^{K'} \\
& \sum_{i\in[K]} \pi_{ij} = \hat{p}^j \quad \forall j \in [K'] \\
& \sum_{j\in[K']} \pi_{ij} = \hat{p}^i \quad \forall i \in [K].
\end{aligned}$$

Assume the conditions in Theorem 1 hold and the true GMM mixture size K belongs to $\mathcal{K}$. Let ε be defined as in Theorem 2. Then, centering the ambiguity set (7) at the conditional distribution $\hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}$, where

$$K' \in \operatorname*{arg\,min}_{K'\in\mathcal{K}} \max_{L\in\mathcal{K}} \overline{\mathbb{W}}_2^2(\hat{\mathbb{M}}^L_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}),$$

and setting the radius to

$$\varepsilon' := \varepsilon + \max_{L \in \mathcal{K}} \overline{\mathbb{W}}_2(\hat{\mathbb{M}}^{L}_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}),$$

ensure that $\mathbb{M}^{K}_{\boldsymbol{\xi}|\boldsymbol{s}} \in \mathcal{P}_{\varepsilon'}$*.*

Proposition 3 offers a theoretically sound method for making decisions that are robust not only to data-driven estimation errors but also to the misspecification of the GMM's complexity. To the best of our knowledge, this is the first work to establish a formal coverage guarantee for GMM under misspecification.

The models and reformulations developed so far leverage the analytical tractability of GMMs, enabling the approximation of the conditional distribution through efficient Monte Carlo sampling. This approach is highly effective for single-stage optimization problems. However, when extended to multistage settings, standard sampling-based dynamic programming methods encounter a significant challenge: the size of the resulting problem often grows exponentially with the number of planning horizons T. To overcome this limitation, the next section introduces an alternative approximation scheme that avoids direct sampling, offering a solution method whose complexity scales much more favorably with the planning horizon.

4. Optimization Models Using Observational Data

In this section, we develop optimization models that directly utilize the observational data, bypassing the need for Monte Carlo sampling to approximate the conditional expectations. These models offer significant advantages in multistage settings, particularly when solving stochastic optimization problems using data-driven dynamic programming (Park et al. 2022), where the problem size scales only linearly in the time horizon and data size.

The traditional scheme for multistage stochastic programming is the sample-average approximation (SAA) scheme (Shapiro 2011), which presumes access to the underlying distribution and relies on sequential conditional Monte Carlo sampling to capture the evolution of the stochastic process. While one could, in principle, apply the SAA scheme using samples from the approximate distribution $\hat{\mathbb{M}}$, the resulting problem size grows exponentially with the time horizon, making dynamic programming methods computationally intractable. By contrast, the data-driven scheme of (Park et al. 2022), operates directly on historical trajectories by reweighting the transition probabilities between consecutive data points using kernel regression to capture the process dynamics. We propose to replace these kernel-based weights with Gaussian-mixture-based weights, leveraging a GMM to approximate conditional transitions. This substitution preserves the direct-from-data workflow while mitigating the curse of dimensionality inherent in kernel regression, and enables scalable multistage optimization without resorting to large-scale Monte Carlo sampling.

We begin by describing the solution procedure in a single-stage setting. Let f and $\hat{f}$ be the density functions of $\mathbb{M}$ and $\hat{\mathbb{M}}$, respectively. Given N historical observations $\{(\boldsymbol{s}_n, \boldsymbol{\xi}_n)\}_{n\in[N]}$ drawn from $\mathbb{M}$, we approximate the contextual stochastic optimization problem (2) by

$$\min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{\hat{f}(\boldsymbol{s})N} \sum_{n\in[N]} \ell(\boldsymbol{x},\boldsymbol{\xi}_n)\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n). \tag{11}$$

We now justify this approximation. The conditional expectation in the objective can be rewritten as

$$\begin{aligned}\mathbb{E}_{\mathbb{M}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\big|\tilde{\boldsymbol{s}}=\boldsymbol{s}\right] &= \int \ell(\boldsymbol{x},\boldsymbol{\xi})f(\boldsymbol{s},\boldsymbol{\xi})/f(\boldsymbol{s})\mathrm{d}\boldsymbol{\xi} \\ &= \int \ell(\boldsymbol{x},\boldsymbol{\xi})f(\boldsymbol{s}|\boldsymbol{\xi})f(\boldsymbol{\xi})/f(\boldsymbol{s})\mathrm{d}\boldsymbol{\xi} \\ &= \frac{1}{f(\boldsymbol{s})}\mathbb{E}_{\mathbb{M}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})f(\boldsymbol{s}|\tilde{\boldsymbol{\xi}})\right].\end{aligned}$$

Using the samples $\{\boldsymbol{\xi}_n\}_{n\in[N]}$ and applying sample-average approximation yields:

$$\mathbb{E}_{\mathbb{M}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\big|\tilde{\boldsymbol{s}}=\boldsymbol{s}\right] \approx \frac{1}{N}\sum_{n\in[N]} \ell(\boldsymbol{x},\boldsymbol{\xi}_n)\frac{f(\boldsymbol{s}|\boldsymbol{\xi}_n)}{f(\boldsymbol{s})}.$$

The factor $\frac{f(\boldsymbol{s}|\boldsymbol{\xi}_n)}{f(\boldsymbol{s})} = \frac{f(s,\boldsymbol{\xi}_n)}{f(\boldsymbol{s})f(\boldsymbol{\xi}_n)}$ constitutes a likelihood ratio that quantifies the local dependence between $\boldsymbol{s}$ and $\boldsymbol{\xi}_n$. Values greater than one indicate that the co-occurrence $(\boldsymbol{s},\boldsymbol{\xi}_n)$ is more likely than under independence, while values smaller than one indicate the opposite (Church and Hanks 1990). Lastly, the formulation (11) is obtained by replacing the unknown true density f with the approximation $\hat{f}$. We further remark that the approximation scheme can be naturally generalized to a random vector $(\tilde{\boldsymbol{s}}', \tilde{\boldsymbol{\xi}}')$ governed by an arbitrary distribution $\mathbb{P}$ through a normalizing flow.

4.1. Approximation Quality

We now proceed to establish the approximation quality of our proposed scheme, demonstrating its favorable sample complexity properties. To this end, we introduce the following regularity condition:

(D) There exist constants $\overline{f}, \underline{f} \in \mathbb{R}_{++}$ such that $\overline{f} \geq \hat{f}(\boldsymbol{s}|\boldsymbol{\xi})$ for all $\boldsymbol{\xi}\in\mathbb{R}^R$, and $\underline{f} \leq \hat{f}(\boldsymbol{s})$ with probability $1-\delta$.

This condition is standard in the nonparametric regression literature (Györfi et al. 2006, Kohler et al. 2009, Belkin et al. 2019) and serves to simplify the presentation. Specifically, since $\hat{f}(\boldsymbol{s}|\boldsymbol{\xi})$ is a conditional density of a mixture of Gaussians—and hence itself a mixture of Gaussians—it admits a uniform upper bound. Likewise, since $\hat{f}(\boldsymbol{s})$ is a marginal density of a Gaussian mixture, one can always determine a lower bound $\underline{f}$ given a confidence level δ. Based on these assumptions, the following theorem bounds the difference between the true and empirical conditional expectations.

Theorem 3. *Let $\mathbb{M} = \sum_{k\in[K]} p^k \mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$ be the true Gaussian mixture distribution and $\hat{\mathbb{M}} = \sum_{k\in[K]} \hat{p}^k \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)$ be the estimated one. Suppose that $|p^k - \hat{p}^k| \leq \epsilon_p$, $\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k| \leq \epsilon_{\boldsymbol{\mu}}$, and $\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \leq \epsilon_{\boldsymbol{\Sigma}}$ for all $k \in [K]$. Then, for any $\delta \in (0,1)$, we have*

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{M}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\middle|\tilde{\boldsymbol{s}}=\boldsymbol{s}\right] - \frac{1}{\hat{f}(\boldsymbol{s})N}\sum_{n\in[N]}\ell(\boldsymbol{x},\boldsymbol{\xi}_n)\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)\right| \\
&\leq \overline{\ell}\left[C_{Q+R}\left(\|\boldsymbol{s}\|^2+\beta+\left(\gamma+\tfrac{\beta}{\alpha}(\|\boldsymbol{s}\|+\gamma)\right)^2\right)+C'_{Q+R}\sqrt{\|\boldsymbol{s}\|^2+\beta+\left(\gamma+\tfrac{\beta}{\alpha}(\|\boldsymbol{s}\|+\gamma)\right)^2}+C''_{Q+R}\right] \\
&\quad+\tfrac{\overline{\ell f}}{\underline{f}}\left(C_R\left(\beta+\gamma^2\right)+C'_R\sqrt{\beta+\gamma^2}+C''_R+C_Q\|\boldsymbol{s}\|^2+C'_Q\|\boldsymbol{s}\|+C''_Q\right) \\
&\quad+\tfrac{\overline{\ell f}}{\underline{f}}\sqrt{\tfrac{1}{2N}\log\left(\tfrac{8}{\delta}\right)}.
\end{aligned} \tag{12}$$

with probability at least $1-\delta$. Here, the constants are defined as $C_D = \left(1+\frac{\epsilon_p}{\underline{p}}\right)\frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}$, $C'_D = \left(1+\frac{\epsilon_p}{\underline{p}}\right)\left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha}+\frac{D\epsilon_{\boldsymbol{\Sigma}}\gamma}{\alpha^2}\right)$, and $C''_D = \frac{\epsilon_p}{\underline{p}}+\left(1+\frac{\epsilon_p}{\underline{p}}\right)\left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha}\gamma+\frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}(\gamma^2+\alpha)+o(\epsilon_\mu+\epsilon_{\boldsymbol{\Sigma}})\right)$.

In Theorem 3, the constants C_{Q+R}, C'_{Q+R}, C''_{Q+R}, C_R, C'_R, C''_R depend linearly on ϵ_p, $\epsilon_{\boldsymbol{\mu}}$, $\epsilon_{\boldsymbol{\Sigma}}$, and the final term decays as $O(\frac{1}{N})$; therefore, the proposed approximation enjoys polynomial sample complexity and avoids the curse of dimensionality. In the next section, we apply this approximation framework to address multistage stochastic optimization problems under Markovian uncertainty.

4.2. Applications to Multistage Stochastic Programming

We consider a multistage stochastic optimization problem with Markovian uncertainty over a planning horizon of T stages, given by:

$$\min_{\boldsymbol{x}_1\in\mathcal{X}_1(\boldsymbol{x}_0,\boldsymbol{\xi}_1)}\ell(\boldsymbol{x}_1,\boldsymbol{\xi}_1)+\mathbb{E}\left[\min_{\boldsymbol{x}_2\in\mathcal{X}_2(\boldsymbol{x}_1,\tilde{\boldsymbol{\xi}}_2)}\ell(\boldsymbol{x}_2,\tilde{\boldsymbol{\xi}}_2)+\mathbb{E}\left[\cdots+\mathbb{E}\left[\min_{\boldsymbol{x}_T\in\mathcal{X}_T(\boldsymbol{x}_{T-1},\tilde{\boldsymbol{\xi}}_T)}\ell(\boldsymbol{x}_T,\tilde{\boldsymbol{\xi}}_T)\middle|\tilde{\boldsymbol{\xi}}_{T-1}\right]\cdots\middle|\tilde{\boldsymbol{\xi}}_2\right]\middle|\boldsymbol{\xi}_1\right]. \tag{13}$$

This problem is traditionally solved via dynamic programming by recursively evaluating the cost-to-go functions from the terminal stage $t=T$ backward to the initial stage $t=1$:

$$V_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)=\min_{\boldsymbol{x}_t\in\mathcal{X}_t(\boldsymbol{x}_{t-1},\tilde{\boldsymbol{\xi}}_t)}\boldsymbol{c}_t^\top\boldsymbol{x}_t+\mathcal{V}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_t)\quad\forall\boldsymbol{x}_{t-1}\in\mathcal{X}_{t-1}\ \forall\boldsymbol{\xi}_t\in\mathbb{R}^Q, \tag{14}$$

where

$$\mathcal{V}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_t)=\mathbb{E}\left[V_{t+1}(\boldsymbol{x}_t,\tilde{\boldsymbol{\xi}}_{t+1})\,\middle|\,\boldsymbol{\xi}_t\right] \tag{15}$$

denotes the conditional expectation of the future cost-to-go function at stage $t+1$, given the most recent realization $\boldsymbol{\xi}_t$. We assume $V_{T+1}(\cdot)=0$, which implies that no additional costs beyond the terminal stage T.

In practice, the underlying stochastic process $\tilde{\boldsymbol{\xi}}_{[T]}=(\tilde{\boldsymbol{\xi}}_1,\tilde{\boldsymbol{\xi}}_2,\ldots,\tilde{\boldsymbol{\xi}}_T)\in\mathbb{R}^{Q\times T}$ is typically unknown and one only has access to N i.i.d. sample trajectories $\{\boldsymbol{\xi}_{[T],n}\coloneqq(\boldsymbol{\xi}_{1,n},\ldots,\boldsymbol{\xi}_{T,n})\}_{n\in[N]}$. In (Park et al. 2022), the authors propose to approximate the conditional expectation in a data-driven fashion

$$\hat{\mathcal{V}}_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_t)\approx\sum_{i\in[N]}w_{t+1}(\boldsymbol{\xi}_t,\boldsymbol{\xi}_{t,i})V_{t+1}(\boldsymbol{x}_t,\boldsymbol{\xi}_{t+1,i}), \tag{16}$$

where the weights $w_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t,i})$ are defined through the kernel regression. This scheme avoids costly evaluations over the continuous space by computing the value functions only at the sample points $\{\boldsymbol{\xi}_{t,n}\}_{n\in[N]}$. However, this method suffers from the curse of dimensionality: its suboptimality scales as $\tilde{\mathcal{O}}(T^{\frac{3}{2}}/N^{\frac{2}{Q+4}})$ and thus deteriorates rapidly as the dimension Q increases.

To overcome this limitation, we propose to replace the kernel regression weights in (16) with the Gaussian-mixture-based observational weights in (11), yielding the approximation

$$\hat{\mathcal{V}}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t) \approx \hat{\mathbb{E}}\left[V_{t+1}(\boldsymbol{x}_t, \tilde{\boldsymbol{\xi}}_{t+1}) | \tilde{\boldsymbol{\xi}}_t = \boldsymbol{\xi}_t\right] = \sum_{i\in[N]} \frac{\hat{f}(\boldsymbol{\xi}_t | \boldsymbol{\xi}_{t+1,i})}{\hat{f}(\boldsymbol{\xi}_t)} V_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_{t+1,i}). \tag{17}$$

This leads to the following suboptimality guarantee for the solution $\widehat{\boldsymbol{x}}_1^N$ obtained via our data-driven dynamic programming approach.

Theorem 4. *Assume that for each t, the cost-to-go functions $V_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)$ is L-Lipschitz continuous in $\boldsymbol{x}_{t-1}$ for any fixed $\boldsymbol{\xi}_t$. Furthermore, assume that the feasible regions are nonempty and compact: there exists a constant $\overline{D} \in \mathbb{R}_{++}$ such that $\sup_{\boldsymbol{x}_t, \boldsymbol{x}'_t \in \mathcal{X}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\xi}_t)} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| \leq \overline{D}$ for all $t \in [T]$, $\boldsymbol{x}_{t-1} \in \mathcal{X}_{t-1}(\cdot)$, and $\boldsymbol{\xi}_t \in \Xi_t$. Then for any $\eta \in \mathbb{R}_{++}$, the following bound holds with probability at least $1-\delta$.*

$$\begin{aligned} &\boldsymbol{\ell}(\widehat{\boldsymbol{x}}_1^N, \boldsymbol{\xi}_1) + \mathcal{V}_2(\widehat{\boldsymbol{x}}_1^N, \boldsymbol{\xi}_1) - \left(\min_{\boldsymbol{x}_1 \in \mathcal{X}_1(\boldsymbol{x}_0, \boldsymbol{\xi}_1)} \ell(\boldsymbol{x}_1, \boldsymbol{\xi}_1) + \mathcal{V}_2(\boldsymbol{x}_1, \boldsymbol{\xi}_1) \right) \\ &\qquad \leq 2(T-1)\tau + 4(T-1)L\eta + 2\sum_{t=2}^{T} \frac{\overline{\ell f}^2}{\underline{f}^2} \sqrt{\frac{2}{N} \log\left(\frac{(T+1)O(1)N^{t-2}(\overline{D}/\eta)^{R(t-1)}}{\delta} \right)} \end{aligned}$$

Here,

$$\begin{aligned} \tau := \bar{\ell} \Bigg[& C_{Q+R}\left(\bar{\xi}^2 + \beta + \left(\gamma + \tfrac{\beta}{\alpha}(\bar{\xi}+\gamma)\right)^2\right) + C'_{Q+R}\sqrt{\bar{\xi}^2 + \beta + \left(\gamma + \tfrac{\beta}{\alpha}(\bar{\xi}+\gamma)\right)^2} + C''_{Q+R} \Bigg] \\ & + \tfrac{\overline{\ell f}}{\underline{f}} \left(C_R\left(\beta + \gamma^2\right) + C'_R\sqrt{\beta+\gamma^2} + C''_R + C_Q\bar{\xi}^2 + C'_Q\bar{\xi} + C''_Q \right). \end{aligned} \tag{18}$$

and $\bar{\xi} := 4\sigma\sqrt{R} + 2\sigma\sqrt{2\log\frac{NT}{\delta}} + \gamma.$

This result establishes that the suboptimality of the proposed approximation decays at the rate

$$\begin{aligned} &\mathcal{O}\left(\sum_{t=2}^{T} \sqrt{\frac{2}{N}\log\left(\frac{\mathcal{O}(1)N^{t-2}(\overline{D}/\eta)^{R(t-1)}}{\delta}\right)} \right) \\ =\ &\mathcal{O}\left(\sum_{t=2}^{T} \sqrt{\frac{2}{N}\left(\log\mathcal{O}(1) + (t-2)\log N + R(t-1)\log\frac{\overline{D}}{\eta} - \log\delta\right)} \right) \\ =\ &\tilde{\mathcal{O}}\left(\sum_{t=2}^{T} \sqrt{\frac{t}{N}} \right) = \tilde{\mathcal{O}}\left(\frac{T^{3/2}}{N^{1/2}}\right), \end{aligned}$$

and crucially, it is free from the curse of dimensionality. In contrast, the kernel-based approach yields a slower rate of $\tilde{\mathcal{O}}(T^{3/2}/N^{2/(Q+4)})$ (Park et al. 2022, Theorem 1). To the best of our knowledge, this is the first result of its kind, even for the special case where the underlying stochastic process is a single multivariate Gaussian ($K = 1$), a setting that frequently arises in multistage portfolio optimization. Notably, the standard sample average approximation approach in this setting only achieves a convergence rate of $\tilde{\mathcal{O}}(T^{1/2}/N^{1/(2T)})$, which deteriorates significantly as the planning horizon T increases (Shapiro and Nemirovski 2005).

4.3. Distributionally Robust Optimization Framework

To enhance the reliability of our data-driven approach, we now embed our model within a DRO framework. To this end, we define the *empirical measure* as

$$\hat{\mathbb{Q}} := \sum_{n\in[N]} \frac{\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)}{\sum_{u\in[N]} \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_u)} \delta_{\boldsymbol{\xi}_n},$$

and use it as the center of the Wasserstein ambiguity set defined in (7). Reformulating the resulting DRO problem into a tractable conic program is straightforward by employing Proposition 2. For the sake of brevity, we omit this derivation and focus on the illustration of how to determine an appropriate radius for the Wasserstein ambiguity set.

The practical application and performance guarantees of the DRO model hinge on choosing a radius that ensures the true conditional distribution $\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}$ is contained within the ambiguity set with high probability. To achieve this, we need a valid upper bound on $\mathbb{W}_2(\hat{\mathbb{Q}}, \mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}})$. Let $\hat{\mathbb{M}}^M_{\boldsymbol{\xi}|\boldsymbol{s}} := \frac{1}{M}\sum_{m\in[M]} \delta_{\boldsymbol{\xi}'_m}$ be the empirical measure constructed from M samples of $\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}$. By the triangle inequality:

$$\mathbb{W}_2(\hat{\mathbb{Q}}, \mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}) \le \mathbb{W}_2(\hat{\mathbb{Q}}, \hat{\mathbb{M}}^M_{\boldsymbol{\xi}|\boldsymbol{s}}) + \mathbb{W}_2(\hat{\mathbb{M}}^M_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}) + \mathbb{W}_2(\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}, \mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}).$$

The first term, $\mathbb{W}_2(\hat{\mathbb{Q}}, \hat{\mathbb{M}}^M_{\boldsymbol{\xi}|\boldsymbol{s}})$, corresponds to the square root of the optimal value of the following linear program:

$$\begin{aligned}
\mathbb{W}_2^2(\hat{\mathbb{Q}}, \hat{\mathbb{M}}^M_{\boldsymbol{\xi}|\boldsymbol{s}}) = \min\ & \sum_{i\in[N]}\sum_{j\in[M]} \pi_{ij} \|\boldsymbol{\xi}_i - \boldsymbol{\xi}'_j\|^2 \\
\text{s.t.}\ & \boldsymbol{\pi} \in \Delta^N \times \Delta^M \\
& \sum_{i\in[N]} \pi_{ij} = \frac{1}{M} \quad \forall j \in [M] \\
& \sum_{j\in[M]} \pi_{ij} = \frac{\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)}{\sum_{u\in[N]} \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_u)} \quad \forall i \in [N].
\end{aligned}$$

The second term, $\mathbb{W}_2(\hat{\mathbb{M}}^M_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}})$, can be bounded using concentration results for empirical Wasserstein distances (Fournier and Guillin 2015, Theorem 2). The final term, $\mathbb{W}_2(\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}, \mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}})$, is upper bounded by the result established in Theorem 2.

Remark 3. The proposed DRO framework can be seamlessly embedded into the multistage stochastic programming formulation described in Section 4.2. Specifically, at each stage t, instead of approximating the conditional expectation $\mathcal{V}_{t+1}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ in (15) via a fixed weighted average as in (17), we solve a distributionally robust optimization problem over a Wasserstein ball centered at the empirical measure $\mathbb{Q}$.

5. Experimental Results

In this section, we conduct numerical experiments to evaluate the out-of-sample performance of different decision-making models. We study three important operations management problems: inventory management, portfolio optimization, and the multistage wind energy planning problem. In each experiment, we begin by describing the problem setting, followed by the model training and hyperparameter tuning procedures, and conclude with out-of-sample evaluation results and their interpretation.

5.1. Inventory Management

We consider an inventory management problem, where the objective of the decision maker is to determine an optimal order quantity under uncertain demand. Specifically, when the order quantity exceeds actual demand, the firm incurs a holding cost. Conversely, when the order quantity is too low, a stock-out cost is penalized due to unmet demand. The loss function under order quantity $q \in \mathbb{R}_+$ and realized demand $\xi \in \mathbb{R}_+$ is given by:

$$\ell(q, \xi) = h(q - \xi)_+ + b(\xi - q)_+,$$

where h and b denote the per-unit holding and stock-out costs, respectively. In this experiment, we set these costs to $h = 10$ and $b = 2$. We assume the demand is influenced by a vector of observable side information $\tilde{\boldsymbol{s}} \in \mathbb{R}^Q$, which is available prior to making the ordering decision.

To assess the capabilities of our proposed method, we design a synthetic data-generating process. Specifically, we construct a multimodal distribution by combining a linear and a quadratic functional relationship. The side information $\tilde{\boldsymbol{s}}$ is drawn from a multivariate uniform distribution, i.e., $\tilde{\boldsymbol{s}} \sim \mathbb{U}(-2, 2)$. For a given realization of $\tilde{\boldsymbol{s}}$, the uncertain demand is then generated as

$$\xi = \begin{cases} 0.3\mathbf{e}^\top \boldsymbol{s} + \delta_1 + \mu_1 & \text{with probability } \frac{1}{2}, \\ 5\sum s_i^2 + \delta_2 + \mu_2 & \text{with probability } \frac{1}{2}, \end{cases}$$

where $\tilde{\delta}_1, \tilde{\delta}_2 \sim \mathbb{U}(-2, 2)$ are uniform noise terms while $\mu_1 = 50,\ \mu_2 = 42.5$ is a base demand. This distribution is intentionally complex and cannot be perfectly captured by a GMM. Furthermore, from a data-driven perspective, a decision-maker unaware of the underlying distribution would find it very difficult to specify a correct functional form for learning. This setup, therefore, represents one of the challenging scenarios in contextual optimization.

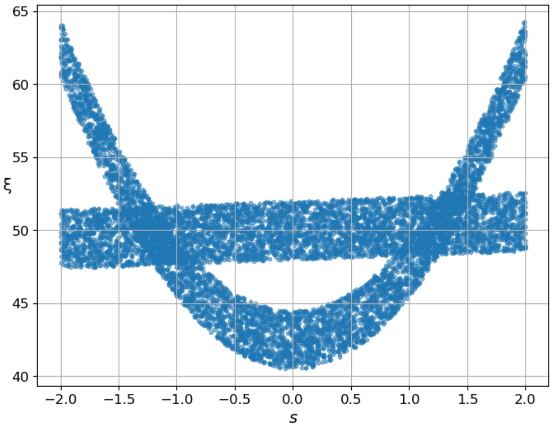


Figure 2 Visualization of the underlying distributions, which is a mixture of quadratic and linear forms.

We evaluate the performance of our proposed models—both the plain-vanilla GMM framework and its extension with normalizing flows—against two prominent benchmarks from the literature. The first benchmark is the regularized Nadaraya-Watson regression (RNW) method (Srivastava et al. 2021). The RNW method is known for enhancing the robustness of kernel-based approaches by incorporating a conditional variance regularization term. It serves as a strong baseline for non-parametric models that do not assume a specific data structure but can be susceptible to the curse of dimensionality. The second benchmark is the residual-based distributionally robust contextual (ResDRO) optimization approach (Kannan et al. 2020b). ResDRO's key strength lies in its flexible residual architecture, which allows it to be effectively paired with various parametric prediction models. This method also employs a DRO framework to improve out-of-sample stability, and it is regarded as a state-of-the-art parametric framework for contextual optimization. Together, these two benchmarks provide a comprehensive comparison, representing the leading non-parametric and parametric approaches in the field.

Table 1 summarizes the out-of-sample performance of all competing methods under three levels of contextual dimensionality: $Q \in \{1, 5, 20\}$. To ensure the statistical significance of our results, the reported metrics are based on 100 independent trials. Within each trial, we train every model on a randomly drawn training set of 100 samples. The performance is then assessed on a large, separately generated test set of 10,000 samples, allowing for a precise estimation of the true out-of-sample conditional expected loss. For reproducibility, the detailed model formulations, hyperparameter tuning procedures, and specifics of the normalizing flow training are provided in Appendix D.1.

The experimental results in Table 1 reveal several key findings. First, one can observe that our GMM-based frameworks demonstrate exceptional performance across all contextual dimensionality. Additionally, the GMM-NF model outperforms the plain-vanilla GMM in higher-dimensional scenarios, demonstrating the effectiveness of the normalizing flow in correcting for model misspecification. Nevertheless, the performance of the original GMM model remains respectable, highlighting its inherent flexibility even under non-GM distributions. In contrast, the parametric model ResDRC struggles under multimodal data-generating distributions. Due to its inability to learn the intricate mixture distribution, this method yields higher out-of-sample loss compared with our GMM models. On the other hand, although we observe that the non-parametric RNW method is competitive at low dimensionality, it suffers a severe performance degradation in higher dimensions due to the curse of dimensionality.

In summary, our proposed GMM framework, particularly when enhanced with normalizing flows, establishes its superiority and robustness across all tested scenarios. It effectively overcomes the curse of dimensionality that plagues the non-parametric benchmark and shows remarkable flexibility compared with parametric models.

Q	model	avg. loss ↓	10^{th} percentile ↓	90^{th} percentile ↓
1	ResDRC	15.406	10.070	22.602
	RNW	11.509	8.869	14.205
	GMM	**9.748**	**5.722**	**12.931**
	GMM-NF	9.880	5.757	12.946
5	ResDRC	31.060	9.459	50.481
	RNW	37.003	15.756	56.568
	GMM	34.986	8.694	61.826
	GMM-NF	**29.518**	**7.549**	**45.351**
20	ResDRC	196.875	113.379	306.892
	RNW	233.464	196.473	275.430
	GMM	143.365	103.331	182.220
	GMM-NF	**140.706**	**94.249**	**174.691**

Table 1 Out-of-sample performances of different contextual optimization methods. GMM represents the plain-vanilla GMM method, and GMM-NF represents the generalized framework with normalizing flows.

5.2. Portfolio Optimization

In the second experiment, we address the problem of contextual portfolio optimization, a critical task in financial decision-making. To ground our analysis in a state-of-the-art context, we adopt the same experimental design as in Nguyen et al. (2024). This alignment ensures that all methods are tested under identical market conditions, facilitating a fair and rigorous comparison. For the sake of completeness, we herein detail the full experimental setup.

Following the work of Chenreddy et al. (2022a), Nguyen et al. (2024), we adopt the following five indices as contextual information: (i) Volatility Index (VIX), (ii) 10-year Treasury Yield Index

(TNX), (iii) Crude Oil Index (CL=F), (iv) S&P 500 (GSPC), (v) Dow Jones Index (DJI). The investment universe consists of 399 assets that constantly stay in S&P 500 from January 01, 2017, to March 31, 2023. The entire dataset is partitioned into three periods: an initial period from January 1, 2017, to December 31, 2018, used for initial model training; a validation period from January 1, 2019, to December 31, 2020, dedicated to hyperparameter tuning; and a final test period from January 1, 2021, to March 31, 2023, for out-of-sample performance evaluation. During validation and out-of-sample test, we employ a rolling-horizon approach, where for any given day t, its training set consists of the preceding two years of data. The investor's objective is to minimize a mean-CVaR with a 10% risk tolerance ($\tau = 0.1$) and a risk-aversion parameter $\eta \in \{1, 3, 5, 7, 9\}$:

$$\mathrm{CVaR}_{\mathbb{P}}^{1-\tau}[-\boldsymbol{x}^\top \tilde{\boldsymbol{\xi}} | \tilde{\boldsymbol{s}} = \boldsymbol{s}] - \eta \mathbb{E}_{\mathbb{P}}[\boldsymbol{x}^\top \tilde{\boldsymbol{\xi}} | \tilde{\boldsymbol{s}} = \boldsymbol{s}]$$

Our proposed methods are benchmarked against a comprehensive suite of established models in portfolio optimization and contextual portfolio optimization literature. These include: (i) the equal-weighted model (EW), (ii) the unconditional Mean-CVaR model (MC), (iii) the Distributionally Robust unconditional Mean-CVaR model (DRMC) (Blanchet and Murthy 2019, Mohajerin Esfahani and Kuhn 2018), (iv) the Conditional Mean-CVaR model (CMC) (Nguyen et al. 2024), (v) the Distributionally Robust Conditional Mean-CVaR model (DRCMC) (Nguyen et al. 2020), (vi) the Optimal Transport (distributional robust) Conditional Mean-CVaR model (OTCMC) (Nguyen et al. 2024), (vii) Data-driven Contextual Optimization with Gaussian Mixture model with normalizing flow (GMM-NF). The implementations of benchmark methods are based on publicly available code,[2] while the hyperparameter tuning procedure and training details of our method are presented in Appendix D.2.

The out-of-sample performance of all methods is summarized in Table 2. The results compellingly demonstrate the superiority of our proposed framework. Specifically, for $\eta = 5$ and 9, our method achieves the best performance, evidenced by a significantly lower out-of-sample loss and a higher Sharpe ratio compared to the benchmark approaches. Meanwhile, for $\eta = 1$, our method also delivers close-to-best performance. This robust outperformance suggests that our framework is adept at modeling the conditional distribution of asset returns, leading to more effective portfolio allocation decisions across all risk preference levels.

5.3. Multi-stage Wind Energy Optimization

In our final experiment, we extend our study to the multistage setting. Following Park et al. (2022) and Kim and Powell (2011), we investigate a multistage energy planning problem in the day-ahead

[2] https://github.com/shanshanwang2019/Robustifying-Conditional-Portfolio-Decisions-via-Optimal-Transport

η	model	risk ↓	mean ↑	CVaR ↓	Sharpe ↑
1	EW	-0.236	0.118	1.501	2.196
	MC	-0.115	0.058	**1.140**	1.366
	DRMC	-0.222	0.111	1.371	**2.301**
	CMC	-0.098	0.049	2.207	0.691
	DRCMC	**-0.237**	0.118	1.498	2.202
	OTCMC	-0.236	0.118	1.492	2.205
	GMM-NF	-0.236	**0.118**	1.476	2.240
3	EW	-0.472	0.118	1.501	2.196
	MC	-0.457	0.114	**1.282**	**2.427**
	DRMC	-0.451	0.113	1.407	2.263
	CMC	-0.051	0.013	2.931	0.129
	DRCMC	-0.473	0.118	1.498	2.202
	OTCMC	-0.469	0.117	1.485	2.208
	GMM-NF	**-0.481**	**0.120**	1.492	2.251
5	EW	-0.709	0.118	1.501	2.196
	MC	-0.670	0.112	**1.412**	2.188
	DRMC	-0.683	0.114	1.432	2.241
	CMC	-0.357	0.059	2.809	0.657
	DRCMC	-0.710	0.118	1.498	2.202
	OTCMC	-0.702	0.117	1.486	2.201
	GMM-NF	**-0.730**	**0.122**	1.503	**2.256**
7	EW	-0.945	0.118	1.501	2.196
	MC	-0.888	0.111	1.598	1.950
	DRMC	-0.917	0.115	**1.448**	2.227
	CMC	-0.725	0.091	3.254	0.832
	DRCMC	-0.946	0.118	1.498	2.202
	OTCMC	-0.937	0.117	1.488	2.199
	GMM-NF	**-0.979**	**0.122**	1.509	**2.257**
9	EW	-1.181	0.118	1.501	2.196
	MC	-1.051	0.105	1.786	1.676
	DRMC	-1.152	0.115	**1.459**	2.217
	CMC	-0.704	0.070	3.845	0.567
	DRCMC	-1.183	0.118	1.498	2.202
	OTCMC	-1.173	0.117	1.490	2.199
	GMM-NF	**-1.228**	**0.123**	1.513	**2.257**

Table 2 Performance metrics for different contextual portfolio optimization models under varying η values.

market. At the beginning of day t, the producer observes the day-ahead hourly prices $\boldsymbol{p}_t \in \mathbb{R}_+^{24}$ and determines the commitment levels $\boldsymbol{u}_t \in \mathbb{R}_+^{24}$ for the next day's production. On day $t+1$, the commitments are fulfilled either through generation $\tilde{\boldsymbol{\xi}}_{t+1}$ or by discharging from the storage units. The unmet commitments incur a penalty of twice the day-ahead price, while the excess generations are used to recharge storage. When the storage capacity is saturated, the surplus energy is curtailed.

We obtain the hourly wind energy data from the North American Land Data Assimilation System (Xia et al. 2013) from 2002 to 2011 at **Ohio** and **North Carolina** regions. The day-ahead prices are publicly available at the PJM market[3]. We construct *weekly trajectories* consisting of seven consecutive days (24 hours × 7 stages), yielding 520 trajectories in total. To capture seasonal variation, we partition the decision-making problem into quarterly subproblems; in each quarter, we take the first three years of data as in-sample trajectories and evaluate out-of-sample performance on the remainder.

We integrate our proposed method into the stochastic dual dynamic programming (SDDP) framework, which is widely considered to be the most effective method for solving multistage stochastic programming problems (Füllner and Rebennack 2025). For the benchmark, we compare our proposed GMM-based SDDP against two popular schemes: stagewise independent SDDP (Independent) (Shapiro 2011), and distributionally robust Nadaraya-Watson SDDP (Nadaraya-Watson) (Park et al. 2022). The first method, independent SDDP, assumes no temporal dependence between different stages. In contrast, the state-of-the-art Nadaraya-Watson method estimates the transition probabilities through Nadaraya-Watson kernel regression and then employs a DRO scheme to robustify the solution. Our proposed approach, however, utilizes Gaussian-mixture-based observational weights to replace the kernel regression weights, as described in (17). Additional experimental details are provided in Appendix D.3.

Data set	model	mean ↑	10^{th} percentile ↑	90^{th} percentile ↑
Ohio	Independent	6.847	2.678	11.673
	Nadaraya-Watson	6.863	3.225	11.254
	GMM	**7.855**	**3.318**	**13.789**
North Carolina	Independent	7.914	0.896	15.341
	Nadaraya-Watson	8.195	1.005	15.642
	GMM	**8.728**	**1.985**	**15.859**

Table 3 Out-of-sample performance of different methods in the multistage day-ahead wind energy planning problem (in $100,000).

Table 3 summarizes the statistics of different methods in the out-of-sample test. Our scheme attains the best results across all the criteria. In contrast, the Nadaraya-Watson SDDP method, while robust, suffers from the curse of dimensionality, which limits its performance in this high-dimensional problem. Our GMM-based approach, however, effectively mitigates this issue by fitting a finite Gaussian mixture to the joint distribution and utilizing its closed-form solution. This allows for a more accurate capture of the temporal relationship, leading to sharper conditional forecasts, better commitment decisions, and ultimately superior out-of-sample performance across all reported metrics.

[3] Day-ahead wind data can be downloaded from PJM: http://dataminer2.pjm.com/feed/da_hrl_lmps/definition

6. Concluding Remarks

This paper introduced a novel framework for data-driven contextual optimization, designed to address a fundamental trade-off in existing solution schemes. Prevailing parametric methods often lack the flexibility to handle complex, multimodal uncertainties, while non-parametric approaches are notoriously hampered by the curse of dimensionality. Our GMM-based framework elegantly bridges this gap, marrying the powerful expressive capabilities of non-parametric models with the favorable sample complexity of their parametric counterparts. Recognizing the practical need for relaxing strict distributional assumptions, we integrated our framework with normalizing flows, thereby generalizing the applicability of our model to arbitrary distributions. Finally, we extend our framework to the domain of dynamic decision-making by designing a novel GMM-based solution scheme for multistage stochastic optimization problems under Markovian uncertainty. Our theoretical analysis confirmed that this method achieves significantly better sample complexity than traditional approaches, providing a powerful new methodology for solving long-horizon, high-dimensional multistage problems. Extensive numerical experiments on a series of financial and business decision problems demonstrate the practical superiority of our approach. The framework consistently outperformed state-of-the-art benchmarks, especially in high-dimensional scenarios and when faced with complex data structures.

References

Hirotugu Akaike. Akaike's information criterion. In *International Encyclopedia of Statistical Science*, pages 41–42. Springer, 2025.

Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. In *Conference on Learning Theory*, pages 1135–1164. PMLR, 2014.

Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes. *Advances in Neural Information Processing Systems*, 31, 2018a.

Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. Sample-efficient learning of mixtures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

William L Beedles. Asymmetry in Australian equity returns. *Australian Journal of Management*, 11(1): 1–12, 1986.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66 (3):1025–1044, 2020.

Dimitris Bertsimas and Nihal Koduri. Data-driven optimization: A reproducing kernel Hilbert space approach. *Operations Research*, 70(1):454–471, 2022.

Dimitris Bertsimas and B. Van Parys. Bootstrap robust prescriptive analytics. *arXiv preprint arXiv:1711.09974*, 2017.

Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru. Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures-Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Harrison John Bhatti and Mike Danilovic. Making the world more sustainable: enabling localized energy generation and distribution on decentralized smart grid systems. *World Journal of Engineering and Technology*, 6(2):350–382, 2018.

John R Birge and Francois Louveaux. *Introduction to Stochastic Programming*. Springer Science & Business Media, 2011.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Alex Botte and Doris Bao. A machine learning approach to regime modeling. *Two Sigma Street Review*, 2021.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Geunyeong Byeon. Comparative analysis of two-stage distributionally robust optimization over 1-Wasserstein and 2-Wasserstein balls. *arXiv preprint arXiv:2501.05619*, 2025.

Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

Arijit Chakrabarti and Jayanta K. Ghosh. AIC, BIC and recent advances in model selection. In *Philosophy of Statistics*, volume 7. North-Holland, 2011.

Li Chen and Erica L Plambeck. Dynamic inventory management with learning about the demand distribution and substitution probability. *Manufacturing & Service Operations Management*, 10(2):236–256, 2008.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018a.

Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

Wenbo Chen, Mathieu Tanneau, and Pascal Van Hentenryck. End-to-end feasible optimization proxies for large-scale economic dispatch. *IEEE Transactions on Power Systems*, 39(2):4723–4734, 2023.

Xi Chen, Zachary Owen, Clark Pixton, and David Simchi-Levi. A statistical learning approach to personalization in revenue management. *Management Science*, 68(3):1923–1937, 2022.

Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018b.

Abhilash Reddy Chenreddy, Nymisha Bandi, and Erick Delage. Data-driven conditional robust optimization. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022a.

Abhilash Reddy Chenreddy, Nymisha Bandi, and Erick Delage. Data-driven conditional robust optimization. *Advances in Neural Information Processing Systems*, 35:9525–9537, 2022b.

Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

Antonio J Conejo, Enrique Castillo, Roberto Mínguez, and Federico Milano. Locational marginal price sensitivities. *IEEE Transactions on Power Systems*, 20(4):2026–2033, 2005.

Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems*, 30, 2017.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in Neural Information Processing Systems*, 32, 2019.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Frank J Fabozzi, Svetlozar T Rachev, and Christian Menn. *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing.* John Wiley & Sons, 2005.

Eugene F Fama and Kenneth R French. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22, 2015.

Evelyn Fix. *Discriminatory Analysis: nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine, 1985.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Christian Füllner and Steffen Rebennack. Stochastic dual dynamic programming and its variants: A review. *SIAM Review*, 67(3):415–539, 2025.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Clark R Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression.* Springer Science & Business Media, 2006.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760, 2015.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pages 2722–2730. PMLR, 2019.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

Michael Huang and Vishal Gupta. Decision-focused learning with directional gradients. *Advances in Neural Information Processing Systems*, 37:79194–79220, 2024.

Jakob Huber, Sebastian Müller, Moritz Fleischmann, and Heiner Stuckenschmidt. A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research*, 278(3):904–915, 2019.

Priyank Jaini, Kira A Selby, and Yaoliang Yu. Sum-of-squares polynomial flow. In *International Conference on Machine Learning*, pages 3009–3018. PMLR, 2019.

Jie Jiang and Shengjie Li. On complexity of multistage stochastic programs under heavy tailed distributions. *Operations Research Letters*, 49(2):265–269, 2021.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling Gaussians. *Communications of the ACM*, 55(2):113–120, 2012.

Olav Kallenberg. *Foundations of Modern Probability*, volume 2. Springer, 1997.

Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Data-driven sample average approximation with covariate information. *Optimization Online. URL: http://www. optimization-online. org/DB_HTML/2020/07/7932. html*, 2020a.

Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Residuals-based distributionally robust optimization with covariate information. *arXiv preprint arXiv:2012.01088*, 2020b.

Leonid Vasilevich Kantorovich and SG Rubinshtein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

Jae Ho Kim and Warren B Powell. Optimal energy commitments with storage and intermittent supply. *Operations Research*, 59(6):1347–1360, 2011.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems*, 29, 2016.

Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

Michael Kohler, Adam Krzyżak, and Harro Walk. Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 139(4):1286–1296, 2009.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

Stanley J Kon. Models of stock returns—a comparison. *The Journal of Finance*, 39(1):147–165, 1984.

Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A flow-based generative model for video. *arXiv preprint arXiv:1903.01434*, 2(5):3, 2019.

Jeongyeol Kwon and Constantine Caramanis. The EM algorithm gives sample-optimality for learning mixtures of well-separated Gaussians. In *Conference on Learning Theory*, pages 2425–2487. PMLR, 2020.

Markus Leippold, Qian Wang, and Wenyu Zhou. Machine learning in the chinese stock market. *Journal of Financial Economics*, 145(2):64–82, 2022.

Bogdan Mazoure, Thang Doan, Audrey Durand, Joelle Pineau, and R Devon Hjelm. Leveraging exploration in off-policy algorithms via normalizing flows. In *Conference on Robot Learning*, pages 430–444. PMLR, 2020.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. Neural importance sampling. *ACM Transactions on Graphics (ToG)*, 38(5):1–19, 2019.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & its Applications*, 9(1):141–142, 1964.

Viet Anh Nguyen, Fan Zhang, Jose Blanchet, Erick Delage, and Yinyu Ye. Distributionally robust local non-parametric conditional estimation. *Advances in Neural Information Processing Systems*, 33:15232–15242, 2020.

Viet Anh Nguyen, Fan Zhang, Shanshan Wang, Jose Blanchet, Erick Delage, and Yinyu Ye. Robustifying conditional portfolio decisions via optimal transport. *Operations Research*, 2024.

Afshin Oroojlooyjadid, Lawrence V Snyder, and Martin Takáč. Applying deep learning to the newsvendor problem. *Iise Transactions*, 52(4):444–463, 2020.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*, 30, 2017.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Hyuk Park, Zhuangzhuang Jia, and Grani A Hanasusanto. Data-driven stochastic dual dynamic programming: Performance guarantees and regularization schemes. *Available at Optimization Online*, 2022.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

John Phillips. On the uniform continuity of operator functions and generalized powers-stormer inequalities. Technical report, 1987.

Meng Qi, Yuanyuan Shi, Yongzhi Qi, Chenxin Ma, Rong Yuan, Di Wu, and Zuo-Jun Shen. A practical end-to-end inventory management model with deep learning. *Management Science*, 69(2):759–773, 2023.

MMCR Reaiche. A note on sample complexity of multistage stochastic programs. *Operations Research Letters*, 44(4):430–435, 2016.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, Inc., 1987.

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.

Huntley Schaller and Simon Van Norden. Regime switching in stock market returns. *Applied Financial Economics*, 7(2):177–191, 1997.

Suvrajeet Sen and Yunxiao Deng. Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. *INFORMS Journal on Optimization (submitted)*, 2018.

A. Shapiro. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10: 353–425, 2003.

Alexander Shapiro. Analysis of stochastic dual dynamic programming method. *European Journal of Operational Research*, 209(1):63–72, 2011.

Alexander Shapiro and Arkadi Nemirovski. On complexity of stochastic programming problems. *Continuous Optimization: Current Trends and Modern Applications*, pages 111–146, 2005.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM, 2021.

Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. Chapman & Hall/CRC, 1986.

Prateek R Srivastava, Purnamrita Sarkar, and Grani A Hanasusanto. A robust spectral clustering algorithm for sub-Gaussian mixture models with outliers. *arXiv preprint arXiv:1912.07546*, 2019.

Prateek R Srivastava, Yijie Wang, Grani A Hanasusanto, and Chin Pang Ho. On data-driven prescriptive analytics with side information: A regularized Nadaraya-Watson approach. *arXiv preprint arXiv:2110.04855*, 2021.

Bo Tang and Elias B Khalil. Pyepo: A pytorch-based end-to-end predict-then-optimize library for linear and integer programming. *Mathematical Programming Computation*, 16(3):297–335, 2024.

Ahmed Touati, Harsh Satija, Joshua Romoff, Joelle Pineau, and Pascal Vincent. Randomized value functions via multiplicative normalizing flows. In *Uncertainty in Artificial Intelligence*, pages 422–432. PMLR, 2020.

Wouter JEC Van Eekelen. A generalized moment approach to sharp bounds for conditional expectations. *arXiv preprint arXiv:2401.00090*, 2023.

Cédric Villani et al. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

Yijie Wang, Ling Dai, Grani A Hanasusanto, and Chin Pang Ho. Robust contextual portfolio optimization with Gaussian mixture models. *Optimization Online*, 2022.

Yijie Wang, Grani A Hanasusanto, and Chin Pang Ho. Generalization bounds for contextual stochastic optimization using kernel regression. *arXiv preprint arXiv:2407.10764*, 2024.

David M Ward. The effect of weather on grid systems and the reliability of electricity supply. *Climatic Change*, 121(1):103–113, 2013.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4): 359–372, 1964.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665, 2019.

Youlong Xia, Brian A Cosgrove, Michael B Ek, Justin Sheffield, Lifeng Luo, Eric F Wood, Kingtse Mo, and NDLAS team. Overview of the north american land data assimilation system(NLDAS). In *Land Surface Observation, Modeling and Data Assimilation*, pages 337–377. World Scientific, 2013.

Chao Zhang, Zihao Zhang, Mihai Cucuringu, and Stefan Zohren. A universal end-to-end approach to portfolio optimization via deep learning. *arXiv preprint arXiv:2111.09170*, 2021.

Chuan Zhang, Yu-Xin Tian, Zhi-Ping Fan, Yang Liu, and Ling-Wei Fan. Product sales forecasting using macroeconomic indicators and online reviews: a method combining prospect theory and sentiment analysis. *Soft Computing*, 24:6213–6226, 2020.

Fuzhen Zhang. *The Schur Complement and Its Applications*, volume 4. Springer Science & Business Media, 2006.

Yanfei Zhang and Junbin Gao. Assessing the performance of deep learning algorithms for newsvendor problem. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part I 24*, pages 912–921. Springer, 2017.

Appendix A: Proofs of Section 2

Proposition 4. *Consider two Gaussians $\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ in $\mathbb{R}^D$. If their total variation distance $\mathbb{TV}(\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}),\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}))$ is less than or equal to ϵ, then*

$$\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\|^2 \lessapprox \frac{16\beta^2}{D} \quad \text{and} \quad \frac{1}{8\beta}\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\| \lessapprox 8\beta.$$

Proof The total variation distance admits a lower bound in Hellinger distance, which in the case of two Gaussians $\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ is given by

$$\mathbb{H}(\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}),\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})) := 1-\frac{|\boldsymbol{\Sigma}|^{\frac{1}{4}}|\hat{\boldsymbol{\Sigma}}|^{\frac{1}{4}}}{|\boldsymbol{\Omega}|^{\frac{1}{2}}}\exp\left(-\frac{1}{8}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\right) \le \mathbb{TV}(\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}),\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})), \tag{19}$$

where $\boldsymbol{\Omega} := \frac{\boldsymbol{\Sigma}+\hat{\boldsymbol{\Sigma}}}{2}$.

We will further derive a lower bound on the Hellinger distance. To simplify the exposition, we denote $\boldsymbol{\Delta} := \boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}$; hence $\boldsymbol{\Sigma}=\boldsymbol{\Omega}+\boldsymbol{\Delta}/2$ and $\hat{\boldsymbol{\Sigma}}=\boldsymbol{\Omega}-\boldsymbol{\Delta}/2$.

First, we derive an upper bound on the ratio $|\boldsymbol{\Sigma}|^{\frac{1}{4}}|\hat{\boldsymbol{\Sigma}}|^{\frac{1}{4}}/|\boldsymbol{\Omega}|^{\frac{1}{2}}$. Applying the logarithm operator to the ratio, we get

$$\log\frac{|\boldsymbol{\Omega}+\boldsymbol{\Delta}/2|^{\frac{1}{4}}\log|\boldsymbol{\Omega}-\boldsymbol{\Delta}/2|^{\frac{1}{4}}}{|\boldsymbol{\Omega}|^{\frac{1}{2}}} = \frac{1}{4}\log|\boldsymbol{\Omega}+\boldsymbol{\Delta}/2|+\frac{1}{4}\log|\boldsymbol{\Omega}-\boldsymbol{\Delta}/2|-\frac{1}{2}\log|\boldsymbol{\Omega}|.$$

The second-order Taylor expansion of log det around $\boldsymbol{\Omega}$ is given by (Boyd and Vandenberghe 2004, Section A.4.3):

$$\begin{aligned}\log|\boldsymbol{\Omega}+\boldsymbol{\Delta}/2| &= \log|\boldsymbol{\Omega}|+\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta})-\frac{1}{8}\mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta})+o(\|\boldsymbol{\Delta}\|^2)\\ \log|\boldsymbol{\Omega}-\boldsymbol{\Delta}/2| &= \log|\boldsymbol{\Omega}|-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta})-\frac{1}{8}\mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta})+o(\|\boldsymbol{\Delta}\|^2).\end{aligned}$$

Hence,

$$\begin{aligned}\log\frac{|\boldsymbol{\Omega}+\boldsymbol{\Delta}/2|^{\frac{1}{4}}|\boldsymbol{\Omega}-\boldsymbol{\Delta}/2|^{\frac{1}{4}}}{|\boldsymbol{\Omega}|^{\frac{1}{2}}} &= -\frac{1}{16}\mathrm{tr}(\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta})+o(\|\boldsymbol{\Delta}\|^2)\\ &= -\frac{1}{16}\|\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-\frac{1}{2}}\|_F^2+o(\|\boldsymbol{\Delta}\|^2).\end{aligned}$$

Applying the exponential operator, we thus obtain

$$\frac{|\boldsymbol{\Sigma}|^{\frac{1}{4}}|\hat{\boldsymbol{\Sigma}}|^{\frac{1}{4}}}{|\boldsymbol{\Omega}|^{\frac{1}{2}}} \le \exp\left(-\frac{1}{16}\|\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-\frac{1}{2}}\|_F^2+o(\|\boldsymbol{\Delta}\|^2)\right).$$

Substituting this bound into the Hellinger distance (19) yields

$$\mathbb{H}(\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}),\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}))$$

$$
\begin{aligned}
&\geq 1-\exp\left(-\frac{1}{16}\|\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-\frac{1}{2}}\|_F^2+o(\|\boldsymbol{\Delta}\|^2)-\frac{1}{8}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\right)\\
&\gtrapprox \frac{1}{16}\|\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-\frac{1}{2}}\|_F^2+\frac{1}{8}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\\
&\geq \frac{1}{16\beta^2}\|\boldsymbol{\Delta}\|_F^2+\frac{1}{8\beta}\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|\\
&\geq \frac{D}{16\beta^2}\|\boldsymbol{\Delta}\|^2+\frac{1}{8\beta}\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|.
\end{aligned}
$$

where we have approximated $\exp(-x)$ around $x=0$ in line 3. Thus, the claim follows. □

Next, we present the proof of Theorem 1. The proof of Theorem 1 relies on several results.

LEMMA 2. *Let $a, a' \in \mathbb{R}$, $b, b' \in \mathbb{R}_+$ and suppose that $b, b' \geq \underline{b} > 0$. If $|a-a'| \leq \epsilon_a$ and $|b-b'| \leq \epsilon_b$ then*

$$\left|\frac{a}{b}-\frac{a'}{b'}\right| \leq \frac{\epsilon_a}{b}+\frac{\epsilon_b|a'|}{bb'} \leq \frac{\epsilon_a}{\underline{b}}+\frac{\epsilon_b|a'|}{\underline{b}^2}.$$

Proof of Lemma 2 We have

$$\left|\frac{a}{b}-\frac{a'}{b'}\right| = \left|\frac{ab'-a'b}{bb'}\right| \leq \frac{|ab'-a'b'|+|a'b'-a'b|}{bb'} \leq \frac{\epsilon_a}{b}+\frac{\epsilon_b|a'|}{bb'} \leq \frac{\epsilon_a}{\underline{b}}+\frac{\epsilon_b|a'|}{\underline{b}^2},$$

which proves the claim. □

LEMMA 3. *Suppose the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are strictly positive definite with $\alpha\boldsymbol{I} \preceq \boldsymbol{A}$, $\alpha\boldsymbol{I} \preceq \boldsymbol{B}$ for $\alpha \in \mathbb{R}_{++}$. If $\|\boldsymbol{B}-\boldsymbol{A}\| \leq \epsilon$, then*

$$\|\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1}\| \leq \frac{\epsilon}{\alpha^2}$$

Proof of Lemma 3. Since $\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1}=\boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{B}^{-1}$, we have

$$\|\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1}\| = \|\boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{B}^{-1}\| \leq \|\boldsymbol{A}^{-1}\|\|(\boldsymbol{B}-\boldsymbol{A})\|\|\boldsymbol{B}^{-1}\| \leq \frac{\epsilon}{\alpha^2}.$$

Thus, the claim follows.

LEMMA 4. *Suppose $\boldsymbol{\mu}, \hat{\boldsymbol{\mu}} \in \mathbb{R}^D$ and $\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}} \in \mathbb{S}_{++}^D$ with $\alpha\mathbb{I} \preceq \boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}$. If*

$$\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\| \leq \epsilon_{\boldsymbol{\mu}}, \|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\| \leq \epsilon_{\boldsymbol{\Sigma}},$$

then for every $\boldsymbol{z} \in \mathbb{R}^D$, we have

$$
\begin{aligned}
&\left|\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})-\mathcal{N}\left(\boldsymbol{z}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right)\right|\\
&\leq \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left(\epsilon_{\boldsymbol{\mu}}\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\|+\frac{D\epsilon_{\boldsymbol{\Sigma}}}{2}\left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|++o(\epsilon_{\boldsymbol{\mu}}+\epsilon_{\boldsymbol{\Sigma}})\right)\\
&\leq \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha}(\|\boldsymbol{z}\|+\gamma)+\frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}\left((\|\boldsymbol{z}\|+\gamma)^2+\alpha\right)+o(\epsilon_{\boldsymbol{\mu}}+\epsilon_{\boldsymbol{\Sigma}})\right).
\end{aligned}
$$

Proof of Lemma 4. Recall that the normal density function is given by

$$\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}.$$

We first construct a first-order approximation around $(\boldsymbol{\mu},\boldsymbol{\Sigma})$. The gradient with respect to the mean vector is

$$\nabla_{\boldsymbol{\mu}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}).$$

To obtain the gradient with respect to the covariance matrix, we first take the logarithm:

$$\log\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) - \log\sqrt{(2\pi)^D} - \frac{1}{2}\log|\boldsymbol{\Sigma}|.$$

Using (Petersen et al. 2008, Equation (63)) to compute the gradient of the quadratic term

$$\nabla_{\boldsymbol{\Sigma}}(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) = -\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}$$

and (Petersen et al. 2008, Equation (57)) to compute the gradient of the logarithm term

$$\nabla_{\boldsymbol{\Sigma}}\log|\boldsymbol{\Sigma}| = \boldsymbol{\Sigma}^{-1},$$

we find that

$$\nabla_{\boldsymbol{\Sigma}}\log\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}.$$

By the chain rule, we have

$$\nabla_{\boldsymbol{\Sigma}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\nabla_{\boldsymbol{\Sigma}}\log\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}\right).$$

Hence, the Taylor expansion around $(\boldsymbol{\mu},\boldsymbol{\Sigma})$,

$$\mathcal{N}\left(\boldsymbol{z}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) - \nabla_{\boldsymbol{\mu}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}) - \operatorname{tr}\left(\nabla_{\boldsymbol{\Sigma}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})(\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}})\right) + R_1(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$$

yields the upper bound

$$\begin{aligned}
&\left|\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) - \mathcal{N}\left(\boldsymbol{z}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right)\right| \\
&\le \left|\nabla_{\boldsymbol{\mu}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})^\top(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\right| + \left|\operatorname{tr}\left(\nabla_{\boldsymbol{\Sigma}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})(\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}})\right)\right| + R_1(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) \\
&\le \epsilon_{\boldsymbol{\mu}}\|\nabla_{\boldsymbol{\mu}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\| + D\epsilon_{\boldsymbol{\Sigma}}\|\nabla_{\boldsymbol{\Sigma}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\| + R_1(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) \\
&= \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left(\epsilon_{\boldsymbol{\mu}}\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\| + \frac{D\epsilon_{\boldsymbol{\Sigma}}}{2}\left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right\|\right) + R_1(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) \\
&\le \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha}\|\boldsymbol{z}-\boldsymbol{\mu}\| + \frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}\left(\|\boldsymbol{z}-\boldsymbol{\mu}\|^2 + \alpha\right)\right) + R_1(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) \\
&\le \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha}(\|\boldsymbol{z}\|+\gamma) + \frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}\left((\|\boldsymbol{x}\|+\gamma)^2 + \alpha\right)\right) + R_1(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}).
\end{aligned}$$

The third line follows from the assumptions $\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\| \le \epsilon_{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\| \le \epsilon_{\boldsymbol{\Sigma}}$, as well as the inequality $\operatorname{tr}(\boldsymbol{AB}) \le D\|\boldsymbol{A}\|\|\boldsymbol{B}\|$ for $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{D\times D}$. Note that the remainder $R_1(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ is bounded by $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,o(\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\| + \|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\|)$. Substituting this bound yields the desired result. □

Proposition 5. *Let $\mathbb{M} = \sum_{k\in[K]} p^k \mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$ and $\hat{\mathbb{M}} = \sum_{k\in[K]} \hat{p}^k \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)$ be two GM distributions in $\mathbb{R}^D$ with densities f and $\hat{f}$, respectively. If $|p^k - \hat{p}^k| \le \epsilon_p$, $\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k| \le \epsilon_{\boldsymbol{\mu}}$, and $\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \le \epsilon_{\boldsymbol{\Sigma}}$ for all $k \in [K]$, then the densities satisfy*

$$|f(\boldsymbol{z}) - \hat{f}(\boldsymbol{z})| \le f(\boldsymbol{z}) \left(C_D\|\boldsymbol{z}\|^2 + C_D'\|\boldsymbol{z}\| + C_D''\right),$$

where $C_D = \left(1 + \frac{\epsilon_p}{\underline{p}}\right) \frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}$, $C_D' = \left(1 + \frac{\epsilon_p}{\underline{p}}\right) \left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha} + \frac{D\epsilon_{\boldsymbol{\Sigma}}\gamma}{\alpha^2}\right)$, *and* $C_D'' = \frac{\epsilon_p}{\underline{p}} + \left(1 + \frac{\epsilon_p}{\underline{p}}\right) \left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha}\gamma + \frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2}(\gamma^2 + \alpha) + o(\epsilon_\mu + \epsilon_{\boldsymbol{\Sigma}})\right)$.

Proof By the triangle inequality, we have

$$\begin{aligned}
&|f(\boldsymbol{z}) - \hat{f}(\boldsymbol{z})| \\
=&\left| \sum_{k\in[K]} p^k \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) - \sum_{k\in[K]} \hat{p}^k \mathcal{N}\left(\boldsymbol{z}|\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k\right) \right| \\
\le&\left| \sum_{k\in[K]} p^k \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) - \sum_{k\in[K]} \hat{p}^k \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) \right| + \left| \sum_{k\in[K]} \hat{p}^k \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) - \sum_{k\in[K]} \hat{p}^k \mathcal{N}\left(\boldsymbol{z}|\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k\right) \right| \\
\le&\sum_{k\in[K]} \epsilon_p \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) + \sum_{k\in[K]} \hat{p}^k \left| \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) - \mathcal{N}\left(\boldsymbol{z}|\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k\right) \right|.
\end{aligned}$$

Since $\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k| \le \epsilon_{\boldsymbol{\mu}}$ and $\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \le \epsilon_{\boldsymbol{\Sigma}}$, Lemma 4 yields:

$$\begin{aligned}
&\left| \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) - \mathcal{N}\left(\boldsymbol{z}|\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k\right) \right| \\
\le&\mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) \left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha} (\|\boldsymbol{z}\| + \gamma) + \frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2} \left((\|\boldsymbol{x}\| + \gamma)^2 + \alpha\right) + o(\epsilon_{\boldsymbol{\mu}} + \epsilon_{\boldsymbol{\Sigma}}) \right).
\end{aligned}$$

We thus obtain

$$\begin{aligned}
&|f(\boldsymbol{z}) - \hat{f}(\boldsymbol{z})| \\
\le&\sum_{k\in[K]} \epsilon_p \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) + \sum_{k\in[K]} \hat{p}^k \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) \left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha} (\|\boldsymbol{z}\| + \gamma) + \frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2} \left((\|\boldsymbol{x}\| + \gamma)^2 + \alpha\right) + o(\epsilon_{\boldsymbol{\mu}} + \epsilon_{\boldsymbol{\Sigma}}) \right) \\
\le&\sum_{k\in[K]} p^k \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k) \left[\frac{\epsilon_p}{\underline{p}} + \left(1 + \frac{\epsilon_p}{\underline{p}}\right) \left(\frac{\epsilon_{\boldsymbol{\mu}}}{\alpha}(\|\boldsymbol{z}\| + \gamma) + \frac{D\epsilon_{\boldsymbol{\Sigma}}}{2\alpha^2} \left((\|\boldsymbol{z}\| + \gamma)^2 + \alpha\right) + o(\epsilon_{\boldsymbol{\mu}} + \epsilon_{\boldsymbol{\Sigma}}) \right) \right],
\end{aligned}$$

where we have upper bounded $\hat{p}^k$ with $p^k + \epsilon_p$ and ϵ_p with $\frac{p^k}{\underline{p}}\epsilon_p$, and rearranged the terms. This completes the proof. □

Proposition 6. *Let $\mathbb{M} = \sum_{k\in[K]} p^k \mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$ and $\hat{\mathbb{M}} = \sum_{k\in[K]} \hat{p}^k \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)$ be two GM distributions in $\mathbb{R}^D$. If $|p^k - \hat{p}^k| \le \epsilon_p$, $\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k| \le \epsilon_{\boldsymbol{\mu}}$, and $\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \le \epsilon_{\boldsymbol{\Sigma}}$ for all $k \in [K]$, then*

$$|p^k_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}}| \le \left(\frac{p^k \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_s, \boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)}{f(\boldsymbol{s})} + \frac{\hat{p}^k \mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_s, \hat{\boldsymbol{\Sigma}}^k_{ss}\right)}{\hat{f}(\boldsymbol{s})} \right) \left(C_Q\|\boldsymbol{s}\|^2 + C_Q'\|\boldsymbol{s}\| + C_Q''\right)$$

$$\begin{aligned}&\leq 2C_Q\|\boldsymbol{s}\|^2+2C_Q'\|\boldsymbol{s}\|+2C_Q'',\\ \|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}-\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\| &\leq \left(\frac{\beta}{\alpha}+1\right)\epsilon_{\boldsymbol{\mu}}+\frac{\alpha+\beta}{\alpha^2}\left(\|\boldsymbol{s}\|+\gamma\right)\epsilon_{\boldsymbol{\Sigma}},\\ \|\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}-\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\| &\leq \left(\frac{\beta}{\alpha}\right)^2\epsilon_{\boldsymbol{\Sigma}}.\end{aligned}$$

Proof of Proposition 6 Note that $f(\boldsymbol{s})=\sum_{k\in[K]}p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)$ and $\hat{f}(\boldsymbol{s})=\sum_{k\in[K]}\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{s}},\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{ss}}\right)$. By Lemma 2,

$$\begin{aligned}|p^k_{\boldsymbol{\xi}|\boldsymbol{s}}-\hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}}| &= \left|\frac{p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{s},\boldsymbol{\Sigma}^k_{ss}\right)}{f(\boldsymbol{s})}-\frac{\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{s},\hat{\boldsymbol{\Sigma}}^k_{ss}\right)}{\hat{f}(\boldsymbol{s})}\right|\\ &\leq \frac{\left|p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)-\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{s}},\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{ss}}\right)\right|}{f(\boldsymbol{s})}+\frac{\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{s},\hat{\boldsymbol{\Sigma}}^k_{ss}\right)\left|f(\boldsymbol{s})-\hat{f}(\boldsymbol{s})\right|}{f(\boldsymbol{s})\hat{f}(\boldsymbol{s})}\\ &\leq \underbrace{\frac{\left|p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)-\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{s}},\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{ss}}\right)\right|}{f(\boldsymbol{s})}}_{\text{(i)}}+\underbrace{\frac{|f(\boldsymbol{s})-\hat{f}(\boldsymbol{s})|}{f(\boldsymbol{s})}}_{\text{(ii)}},\end{aligned}$$

where we use the fact that $\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{s}},\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{ss}}\right)/\hat{f}(\boldsymbol{s})\leq 1$ in the second line. We now bound each term separately:

(i) Using the same derivation as in the proof of Proposition 5, we find that

$$\left|p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)-\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{s}},\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{ss}}\right)\right|\leq p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{s},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)\left(C_Q\|\boldsymbol{s}\|^2+C_Q'\|\boldsymbol{s}\|+C_Q''\right).$$

Hence,

$$\begin{aligned}\frac{\left|p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)-\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{s}},\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{ss}}\right)\right|}{f(\boldsymbol{s})} &\leq \frac{\left|p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)-\hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{s}},\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{ss}}\right)\right|}{p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{s},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)}\\ &\leq C_Q\|\boldsymbol{s}\|^2+C_Q'\|\boldsymbol{s}\|+C_Q'',\end{aligned}$$

where the first inequality follows from $f(\boldsymbol{s})\geq p^k\mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_{\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)$.

(ii) By Proposition 5, we have

$$\frac{\left|f(\boldsymbol{s})-\hat{f}(\boldsymbol{s})\right|}{f(\boldsymbol{s})}\leq C_Q\|\boldsymbol{s}\|^2+C_Q'\|\boldsymbol{s}\|+C_Q''.$$

Thus, combining both bounds, we obtain

$$|p^k_{\boldsymbol{\xi}|\boldsymbol{s}}-\hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}}|\leq 2C_Q\|\boldsymbol{s}\|^2+2C_Q'\|\boldsymbol{s}\|+2C_Q''.$$

The error bounds for the mean and covariance estimates are derived in (Wang et al. 2022, Lemma 9):

$$\|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}-\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\|\leq\left(\frac{\beta}{\alpha}+1\right)\epsilon_{\boldsymbol{\mu}}+\frac{\alpha+\beta}{\alpha^2}\left(\|\boldsymbol{s}\|+\gamma\right)\epsilon_{\boldsymbol{\Sigma}},$$

$$\|\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\| \le \left(\frac{\beta}{\alpha}\right)^2 \epsilon_{\boldsymbol{\Sigma}}.$$

Thus, the claim follows. □

LEMMA 5. *If $\tilde{\boldsymbol{z}} \sim \mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a Gaussian vector in $\mathbb{R}^D$ then $\mathbb{E}_{\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\|\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})\|] \le \sqrt{D}\|\boldsymbol{\Sigma}^{-1}\|$.*

Proof of Lemma 5 Let $\tilde{\boldsymbol{z}} \sim \mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Then, $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\tilde{\boldsymbol{z}}-\boldsymbol{\mu}) \sim \mathbb{N}(\boldsymbol{0},\mathbb{I})$, and thus $\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}-\boldsymbol{\mu}) \sim \mathbb{N}(\boldsymbol{0},\boldsymbol{\Sigma}^{-1})$. We have

$$\mathbb{E}_{\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\|\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})\|] = \mathbb{E}_{\mathbb{N}(\boldsymbol{0},\boldsymbol{\Sigma}^{-1})}[\|\tilde{\boldsymbol{z}}\|] \le \sqrt{\mathbb{E}_{\mathbb{N}(\boldsymbol{0},\boldsymbol{\Sigma}^{-1})}[\|\tilde{\boldsymbol{z}}\|^2]} = \sqrt{\mathrm{tr}(\boldsymbol{\Sigma}^{-1})} \le \sqrt{D}\|\boldsymbol{\Sigma}^{-1}\|.$$

where we use Jensen's inequality to upper bound the expectation of a norm. This completes the proof. □

LEMMA 6. *If $\tilde{\boldsymbol{z}} \sim \mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a Gaussian vector in $\mathbb{R}^D$ then*

$$\mathbb{E}_{\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\left\|\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right\|\right] \le D\|\boldsymbol{\Sigma}^{-1}\|.$$

Proof of Lemma 6 Let $\tilde{\boldsymbol{z}} \sim \mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Then, $\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}-\boldsymbol{\mu}) \sim \mathbb{N}(\boldsymbol{0},\boldsymbol{\Sigma}^{-1})$. By (Koltchinskii and Lounici 2017, Theorem 4) we have:

$$\begin{aligned}
\mathbb{E}_{\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\left\|\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right\|\right] &\le \|\boldsymbol{\Sigma}^{-1}\| \max\left\{\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma}^{-1})}{\|\boldsymbol{\Sigma}^{-1}\|}}, \frac{\mathrm{tr}(\boldsymbol{\Sigma}^{-1})}{\|\boldsymbol{\Sigma}^{-1}\|}\right\} \\
&= \max\left\{\sqrt{\|\boldsymbol{\Sigma}^{-1}\|\mathrm{tr}(\boldsymbol{\Sigma}^{-1})}, \mathrm{tr}(\boldsymbol{\Sigma}^{-1})\right\} \\
&\le \max\left\{\sqrt{D}\|\boldsymbol{\Sigma}^{-1}\|, D\|\boldsymbol{\Sigma}^{-1}\|\right\} \\
&= D\|\boldsymbol{\Sigma}^{-1}\|.
\end{aligned}$$

Thus, the claim follows. □

Proof of Theorem 1 Let f and $\hat{f}$ be the density functions of $\mathbb{M}$ and $\hat{\mathbb{M}}$, respectively. We have

$$\begin{aligned}
\left|\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] - \mathbb{E}_{\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]\right| &= \left|\int \ell(\boldsymbol{x},\boldsymbol{\xi}) f(\boldsymbol{\xi}|\boldsymbol{s})\mathrm{d}\boldsymbol{\xi} - \int \ell(\boldsymbol{x},\boldsymbol{\xi})\hat{f}(\boldsymbol{\xi}|\boldsymbol{s})\mathrm{d}\boldsymbol{\xi}\right| \\
&= \left|\int \ell(\boldsymbol{x},\boldsymbol{\xi})\left(f(\boldsymbol{\xi}|\boldsymbol{s}) - \hat{f}(\boldsymbol{\xi}|\boldsymbol{s})\right)\mathrm{d}\boldsymbol{\xi}\right| \\
&\le \overline{\ell}\int \left|f(\boldsymbol{\xi}|\boldsymbol{s}) - \hat{f}(\boldsymbol{\xi}|\boldsymbol{s})\right|\mathrm{d}\boldsymbol{\xi}.
\end{aligned}$$

We next bound the difference of conditional densities:

$$\left|f(\boldsymbol{\xi}|\boldsymbol{s}) - \hat{f}(\boldsymbol{\xi}|\boldsymbol{s})\right|$$

$$
\begin{aligned}
&= \left| \sum_{k\in[K]} p^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) - \sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathcal{N}\left(\boldsymbol{\xi}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \right| \\
&\le \left| \sum_{k\in[K]} p^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) - \sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \right| \\
&\qquad + \left| \sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) - \sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathcal{N}\left(\boldsymbol{\xi}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \right| \\
&\le \sum_{k\in[K]} \left| p^k_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \right| \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) + \sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \left| \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) - \mathcal{N}\left(\boldsymbol{\xi}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \right|.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\left| \mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] - \mathbb{E}_{\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \right| \\
\le& \underbrace{\bar{\ell} \int \sum_{k\in[K]} \left| p^k_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \right| \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \mathrm{d}\boldsymbol{\xi}}_{\text{(i)}} + \underbrace{\bar{\ell} \int \sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \left| \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) - \mathcal{N}\left(\boldsymbol{\xi}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \right| \mathrm{d}\boldsymbol{\xi}}_{\text{(ii)}}.
\end{aligned}
$$

We analyze each integral separately:

(i) Using Proposition 6, we obtain

$$
\begin{aligned}
&\int \sum_{k\in[K]} \left| p^k_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \right| \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \mathrm{d}\boldsymbol{\xi} \\
\le\ & \left(C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right) \sum_{k\in[K]} \left(\frac{p^k \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_s, \boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)}{f(\boldsymbol{s})} + \frac{\hat{p}^k \mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_s, \hat{\boldsymbol{\Sigma}}^k_{ss}\right)}{\hat{f}(\boldsymbol{s})} \right) \int \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \mathrm{d}\boldsymbol{\xi} \\
=\ & 2\left(C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right).
\end{aligned}
$$

(ii) Using Lemma 4, we have

$$
\begin{aligned}
&\left| \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) - \mathcal{N}\left(\boldsymbol{\xi}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \right| \\
\le\ & \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \Bigg(\epsilon'_{\boldsymbol{\mu}} \|(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}(\boldsymbol{\xi} - \boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})\| \\
&\qquad + \frac{R\epsilon'_{\boldsymbol{\Sigma}}}{2} \left\| (\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}(\boldsymbol{\xi} - \boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})(\boldsymbol{\xi} - \boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^\top (\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1} - (\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1} \right\| + o(\epsilon_{\boldsymbol{\mu}} + \epsilon_{\boldsymbol{\Sigma}}) \Bigg),
\end{aligned}
$$

where $\epsilon'_{\boldsymbol{\mu}} = \left(\frac{\beta}{\alpha} + 1\right)\epsilon_{\boldsymbol{\mu}} + \frac{\alpha+\beta}{\alpha^2}\left(\|\boldsymbol{s}\| + \gamma\right)\epsilon_{\boldsymbol{\Sigma}}$ and $\epsilon'_{\boldsymbol{\Sigma}} = \left(\frac{\beta}{\alpha}\right)^2 \epsilon_{\boldsymbol{\Sigma}}$ by Proposition 6. Hence,

$$
\begin{aligned}
&\int \sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \left| \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) - \mathcal{N}\left(\boldsymbol{\xi}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right) \right| \mathrm{d}\boldsymbol{\xi} \\
\le\ & \sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathbb{E}_{\mathbb{N}(\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})} \left[\epsilon'_{\boldsymbol{\mu}} \|(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}(\tilde{\boldsymbol{\xi}} - \boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})\| \right] + o(\epsilon_{\boldsymbol{\mu}} + \epsilon_{\boldsymbol{\Sigma}})
\end{aligned}
$$

$$+\sum_{k\in[K]} \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathbb{E}_{\mathbb{N}(\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})}\left[\frac{R\epsilon'_{\boldsymbol{\Sigma}}}{2}\left\|(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}(\tilde{\boldsymbol{\xi}}-\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})(\tilde{\boldsymbol{\xi}}-\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^\top(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}-(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}\right\|\right].$$

By Lemma 5, we have

$$\mathbb{E}_{\mathbb{N}(\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})}\left[\|(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}(\tilde{\boldsymbol{\xi}}-\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})\|\right]\le\sqrt{R}\left\|(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}\right\|,$$

and by Lemma 6, we have

$$\mathbb{E}_{\mathbb{N}(\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})}\left[\left\|(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}(\tilde{\boldsymbol{\xi}}-\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})(\tilde{\boldsymbol{\xi}}-\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^\top(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}-(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}\right\|\right]\le R\left\|(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}\right\|.$$

In addition, since $\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}} = \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}\boldsymbol{\xi}} - \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}\boldsymbol{s}}(\boldsymbol{\Sigma}^k_{\boldsymbol{s}\boldsymbol{s}})^{-1}\boldsymbol{\Sigma}^k_{\boldsymbol{s}\boldsymbol{\xi}}$ constitutes a Schur complement of $\boldsymbol{\Sigma}^k$, we must have $\lambda_{\max}\left((\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}\right)\le\lambda_{\max}((\boldsymbol{\Sigma}^k)^{-1})\le\frac{1}{\alpha}$ (Zhang 2006, Lemma 2.3). This yields:

$$\begin{aligned}
&\int\sum_{k\in[K]}\hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\left|\mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}},\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right)-\mathcal{N}\left(\boldsymbol{\xi}|\hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}},\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\right)\right|\mathrm{d}\boldsymbol{\xi}\\
\le&\sum_{k\in[K]}\hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\left(\sqrt{R}\epsilon'_{\boldsymbol{\mu}}+\frac{1}{2}R^2\epsilon'_{\boldsymbol{\Sigma}}\right)\left\|(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})^{-1}\right\|+o(\epsilon_{\boldsymbol{\mu}}+\epsilon_{\boldsymbol{\Sigma}})\\
\le&\sum_{k\in[K]}\hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\left(\sqrt{R}\epsilon'_{\boldsymbol{\mu}}+\frac{1}{2}R^2\epsilon'_{\boldsymbol{\Sigma}}\right)\frac{1}{\alpha}+o(\epsilon_{\boldsymbol{\mu}}+\epsilon_{\boldsymbol{\Sigma}})\\
=&\left(\sqrt{R}\epsilon'_{\boldsymbol{\mu}}+\frac{1}{2}R^2\epsilon'_{\boldsymbol{\Sigma}}\right)\frac{1}{\alpha}+o(\epsilon_{\boldsymbol{\mu}}+\epsilon_{\boldsymbol{\Sigma}}).
\end{aligned}$$

Combining both bounds, we obtain

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]-\mathbb{E}_{\hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]\right|\\
\le&\bar{\ell}\left(2\left(C_Q\|\boldsymbol{s}\|^2+C'_Q\|\boldsymbol{s}\|+C''_Q\right)+\left(\sqrt{R}\epsilon'_{\boldsymbol{\mu}}+\frac{1}{2}R^2\epsilon'_{\boldsymbol{\Sigma}}\right)\frac{1}{\alpha}\right),
\end{aligned}$$

which proves the claim. □

Proof of Proposition 1 The conditional density of $\tilde{\boldsymbol{\xi}}'|\boldsymbol{s}'$ is given by

$$f_{\mathbb{P}}(\boldsymbol{\xi}'|\boldsymbol{s}')=\frac{f_{\mathbb{P}}(\boldsymbol{s}',\boldsymbol{\xi}')}{f_{\mathbb{P}}(\boldsymbol{s}')}.$$

Applying the change-of-variables formula to both the numerator and the denominator yields:

$$f_{\mathbb{P}}(\boldsymbol{\xi}'|\boldsymbol{s}')=\frac{f_{\mathbb{M}}\left(T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}',\boldsymbol{\xi}')\right))\left|\det\mathbf{J}_{T_{\boldsymbol{\theta}}^{-1}}(\boldsymbol{s}',\boldsymbol{\xi}')\right|}{f_{\mathbb{M}}\left(T_{\boldsymbol{\theta},\boldsymbol{s}'}^{-1}(\boldsymbol{s}')\right)\left|\det\mathbf{J}_{T_{\boldsymbol{\theta},\boldsymbol{s}'}^{-1}}(\boldsymbol{s}')\right|}.$$

Since $T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}',\boldsymbol{\xi}')=\left(\boldsymbol{s},T_{\boldsymbol{\theta},\boldsymbol{\xi}'}^{-1}\left(\boldsymbol{s},\boldsymbol{\xi}'\right)\right)$, the ratio of densities reduces to the conditional density:

$$\frac{f_{\mathbb{M}}\left(T_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}',\boldsymbol{\xi}')\right)}{f_{\mathbb{M}}\left(T_{\boldsymbol{\theta},\boldsymbol{s}'}^{-1}(\boldsymbol{s}')\right)}=\frac{f_{\mathbb{M}}\left(\boldsymbol{s},T_{\boldsymbol{\theta},\boldsymbol{\xi}'}^{-1}\left(\boldsymbol{s},\boldsymbol{\xi}'\right)\right)}{f_{\mathbb{M}}\left(\boldsymbol{s}\right)}=f_{\mathbb{M}}\left(T_{\boldsymbol{\theta},\boldsymbol{\xi}'}^{-1}(\boldsymbol{s},\boldsymbol{\xi}')\middle|\boldsymbol{s}\right).\qquad(20)$$

Next, using the block-lower-triangular structure of the Jacobian:

$$\mathbf{J}_{T_{\boldsymbol{\theta}}^{-1}}(\boldsymbol{s}',\boldsymbol{\xi}') = \begin{bmatrix} \dfrac{\partial T^{-1}_{\boldsymbol{\theta},\boldsymbol{s}'}(\boldsymbol{s}')}{\partial \boldsymbol{s}'} & \mathbf{0} \\ \dfrac{\partial T^{-1}_{\boldsymbol{\theta},\boldsymbol{\xi}'}(T^{-1}(\boldsymbol{s}'),\boldsymbol{\xi}')}{\partial \boldsymbol{s}'} & \dfrac{\partial T^{-1}_{\boldsymbol{\theta},\boldsymbol{\xi}'}(\boldsymbol{s},\boldsymbol{\xi}')}{\partial \boldsymbol{\xi}'} \end{bmatrix}$$

we have

$$\begin{aligned} \det \mathbf{J}_{T_{\boldsymbol{\theta}}^{-1}}(\boldsymbol{s}',\boldsymbol{\xi}') &= \det \frac{\partial T^{-1}_{\boldsymbol{\theta},\boldsymbol{s}'}(\boldsymbol{s}')}{\partial \boldsymbol{s}'} \cdot \det \frac{\partial T^{-1}_{\boldsymbol{\theta},\boldsymbol{\xi}'}(\boldsymbol{s},\boldsymbol{\xi}')}{\partial \boldsymbol{\xi}'} \\ &= \det \mathbf{J}_{T^{-1}_{\boldsymbol{\theta},\boldsymbol{s}'}}(\boldsymbol{s}') \cdot \det \mathbf{J}_{T^{-1}_{\boldsymbol{\theta},\boldsymbol{\xi}'}}(\boldsymbol{s},\boldsymbol{\xi}') \end{aligned}$$

Hence, the ratio of determinants simplifies to

$$\frac{\left|\det \mathbf{J}_{T_{\boldsymbol{\theta}}^{-1}}(\boldsymbol{s}',\boldsymbol{\xi}')\right|}{\left|\det \mathbf{J}_{T^{-1}_{\boldsymbol{\theta},\boldsymbol{s}'}}(\boldsymbol{s}')\right|} = \left|\det \mathbf{J}_{T^{-1}_{\boldsymbol{\theta},\boldsymbol{\xi}'}}(\boldsymbol{s},\boldsymbol{\xi}')\right|. \tag{21}$$

Combining (20) and (21) yields the desired result. □

Appendix B: Proofs of Section 3

Proof of Proposition 2 Under the prescribed piecewise affine loss function, the formulation (10) becomes

$$\min_{\boldsymbol{x}\in\mathcal{X},\lambda\in\mathbb{R}_+} \varepsilon^2\lambda + \frac{1}{M}\sum_{m\in[M]} \sup_{\boldsymbol{\omega}\in\mathbb{R}^R} \max_{j\in[J]} \boldsymbol{a}_j(\boldsymbol{x})^\top\boldsymbol{\omega} + b_j(\boldsymbol{x}) - \lambda\|\boldsymbol{\omega}-\boldsymbol{\xi}_m\|^2.$$

Introducing epigraphical variables $\boldsymbol{\gamma}\in\mathbb{R}^M$ to bring the summands into the constraint system, we get:

$$\begin{array}{ll} \min & \varepsilon^2\lambda + \dfrac{1}{M}\displaystyle\sum_{m\in[M]} \gamma_m \\ \text{s.t.} & \boldsymbol{x}\in\mathcal{X},\ \lambda\in\mathbb{R}_+,\ \boldsymbol{\gamma}\in\mathbb{R}^M \\ & \boldsymbol{a}_j(\boldsymbol{x})^\top\boldsymbol{\omega} + b_j(\boldsymbol{x}) - \lambda\|\boldsymbol{\omega}-\boldsymbol{\xi}_m\|^2 \le \gamma_m \quad \forall\boldsymbol{\omega}\in\mathbb{R}^R \quad \forall j\in[J] \quad \forall m\in[M] \end{array} \tag{22}$$

Each semi-infinite constraint is equivalent to a semidefinite constraint:

$$\begin{aligned} & \boldsymbol{a}_j(\boldsymbol{x})^\top\boldsymbol{\xi} + b_j(\boldsymbol{x}) - \lambda\|\boldsymbol{\omega}-\boldsymbol{\xi}_m\|^2 \le \gamma_m \quad \forall\boldsymbol{\omega}\in\mathbb{R}^R \\ \iff & 0 \le -\boldsymbol{a}_j(\boldsymbol{x})^\top\boldsymbol{\omega} - b_j(\boldsymbol{x}) + \lambda\|\boldsymbol{\omega}-\boldsymbol{\xi}_m\|^2 + \gamma_m \quad \forall\boldsymbol{\omega}\in\mathbb{R}^R \\ \iff & 0 \le -\boldsymbol{a}_j(\boldsymbol{x})^\top\boldsymbol{\omega} - b_j(\boldsymbol{x}) + \lambda\boldsymbol{\omega}^\top\boldsymbol{\omega} - 2\lambda\boldsymbol{\xi}_m^\top\boldsymbol{\omega} + \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m \quad \forall\boldsymbol{\omega}\in\mathbb{R}^R \\ \iff & \mathbf{0} \preceq \begin{bmatrix} \lambda\mathbb{I} & -\left(\lambda\boldsymbol{\xi}_m + \frac{\boldsymbol{a}_j(\boldsymbol{x})}{2}\right) \\ -\left(\lambda\boldsymbol{\xi}_m + \frac{\boldsymbol{a}_j(\boldsymbol{x})}{2}\right)^\top & \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m - b_j(\boldsymbol{x}) \end{bmatrix}. \end{aligned}$$

By the Schur complement, this constraint is equivalent to the hyperbolic constraint:

$$\lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m - b_j(\boldsymbol{x}) \ge \frac{\left(\lambda\boldsymbol{\xi}_m + \frac{\boldsymbol{a}_j(\boldsymbol{x})}{2}\right)^\top\left(\lambda\boldsymbol{\xi}_m + \frac{\boldsymbol{a}_j(\boldsymbol{x})}{2}\right)}{\lambda}, \quad \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m - b_j(\boldsymbol{x}) \ge 0$$

$$\iff \left\| \begin{bmatrix} 2\lambda\boldsymbol{\xi}_m + \boldsymbol{a}_j(\boldsymbol{x}) \\ \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m - b_j(\boldsymbol{x}) - \lambda \end{bmatrix} \right\| \leq \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m - b_j(\boldsymbol{x}) + \lambda, \quad \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \gamma_m - b_j(\boldsymbol{x}) \geq 0.$$

Substituting this constraint back into (22) yields the desired formulation. □

We then present the proof of Theorem 2. To this end, we first derive several auxiliary results.

Proposition 7. *The 2-Wasserstein distance between two Gaussians* $\mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ *in* $\mathbb{R}^D$ *admits the following upper bound:*

$$\mathbb{W}_2^2(\mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 + D\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|.$$

Proof of Propostion 7 The 2-Wasserstein distance between two Gaussians $\mathbb{N}$ and $\hat{\mathbb{N}}$ (Givens and Shortt 1984, Proposition 7) has the closed form formula given by

$$\mathbb{W}_2^2(\mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}} - 2((\boldsymbol{\Sigma}^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}\right).$$

It remains to upper bound the trace term. From (Bhatia et al. 2019, Theorem 1), we have $\operatorname{tr}\left(\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}} - 2((\boldsymbol{\Sigma}^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}\right)^{\frac{1}{2}} = \min_{\boldsymbol{U} \in \mathcal{U}(d)} \|\boldsymbol{\Sigma}^{\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{U}\|_F$, where $\mathcal{U}(D)$ is the set of all $D \times D$ unitary matrices. Setting $\boldsymbol{U} = \mathbb{I}$, we get

$$\operatorname{tr}\left(\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}} - 2((\boldsymbol{\Sigma}^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}\right)^2 \leq \|\boldsymbol{\Sigma}^{\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\|_F^2 \leq D\|\boldsymbol{\Sigma}^{\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\|^2.$$

Since $\|\boldsymbol{\Sigma}^{\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\|^2 \leq \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|$ (Phillips 1987), we thus obtain

$$\operatorname{tr}\left(\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}} - 2((\boldsymbol{\Sigma}^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}\right)^2 \leq D\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|.$$

Combining the bounds yields the desired result. □

Proposition 8. *Let* $\mathbb{M} = \sum_{k \in [K]} p^k \mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$ *and* $\hat{\mathbb{M}} = \sum_{k \in [K]} \hat{p}^k \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)$ *be two GM distributions in* $\mathbb{R}^D$. *If* $\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k| \leq \epsilon_{\boldsymbol{\mu}}$ *and* $\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \leq \epsilon_{\boldsymbol{\Sigma}}$ *for all* $k \in [K]$, *then*

$$\mathbb{W}_2^2(\mathbb{M}, \hat{\mathbb{M}}) \leq \epsilon_{\boldsymbol{\mu}}^2 + D\epsilon_{\boldsymbol{\Sigma}} + \left(4\gamma^2 + 2D\beta\right) \sum_{i \in [K]} |p^i - \hat{p}^i|.$$

Proof of Proposition 8 From (Delon and Desolneux 2020, Section 4) and (Chen et al. 2018b, Section 3), we find that for any two GM distributions $\mathbb{M}$ and $\hat{\mathbb{M}}$ their 2-Wasserstein distance is upper bounded by:

$$\begin{aligned} \mathbb{W}_2^2(\mathbb{M}, \hat{\mathbb{M}}) \leq \min \ & \sum_{i,j \in [K]} \pi_{ij} \mathbb{W}_2^2(\mathbb{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^j, \hat{\boldsymbol{\Sigma}}^j)) \\ \text{s.t.} \ & \boldsymbol{\pi} \in \Delta^K \times \Delta^K \\ & \sum_{i \in [K]} \pi_{ij} = \hat{p}^j \quad \forall j \in [K] \\ & \sum_{j \in [K]} \pi_{ij} = p^i \quad \forall i \in [K]. \end{aligned} \tag{23}$$

Suppose that $p^k = \hat{p}^k + \delta_k$ for $\delta_k \in [-1,1]$. By construction, we must have $\sum_{k\in[K]} \delta_k = 0$ in order for both $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$ to be probability mass functions. Consider now the solution

$$\pi_{ij} = \begin{cases} \hat{p}^i + \frac{\delta_i}{K} & \text{if } i = j \\ \frac{\delta_i}{K} & \text{if } i \neq j. \end{cases}$$

One can verify that this solution is feasible to (23), which implies that

$$\begin{aligned} \mathbb{W}_2^2(\mathbb{M}, \hat{\mathbb{M}}) &\le \sum_{k\in[K]} \left(\hat{p}^k + \frac{\delta_k}{K}\right) \mathbb{W}_2^2(\mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)) + \sum_{i,j\in[K]:i\neq j} \frac{\delta_i}{K} \mathbb{W}_2^2(\mathbb{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^j, \hat{\boldsymbol{\Sigma}}^j)) \\ &= \sum_{k\in[K]} \hat{p}^k \mathbb{W}_2^2(\mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k)) + \sum_{i,j\in[K]} \frac{\delta_i}{K} \mathbb{W}_2^2(\mathbb{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^j, \hat{\boldsymbol{\Sigma}}^j)). \end{aligned}$$

Using Proposition 7 and by our assumptions, we thus arrive at

$$\begin{aligned} \mathbb{W}_2^2(\mathbb{M}, \hat{\mathbb{M}}) &\le \sum_{k\in[K]} \hat{p}^k \left(\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k\|^2 + D\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\|\right) + \sum_{i,j\in[K]} \frac{\delta_i}{K} \left(\|\boldsymbol{\mu}^i - \hat{\boldsymbol{\mu}}^j\|^2 + D\|\boldsymbol{\Sigma}^i - \hat{\boldsymbol{\Sigma}}^j\|\right) \\ &\le \sum_{k\in[K]} \hat{p}^k \left(\epsilon_{\boldsymbol{\mu}}^2 + D\epsilon_{\boldsymbol{\Sigma}}\right) + \sum_{i,j\in[K]} \frac{|\delta_i|}{K} \left(4\gamma^2 + 2D\beta\right) \\ &= \sum_{k\in[K]} \hat{p}^k \left(\epsilon_{\boldsymbol{\mu}}^2 + D\epsilon_{\boldsymbol{\Sigma}}\right) + \left(4\gamma^2 + 2D\beta\right) \sum_{i\in[K]} |\delta_i|. \end{aligned}$$

This completes the proof. □

Equipped with these preliminary results, we now present the proof of Theorem 2.

Proof of Theorem 2 By Proposition 6, we have

$$\begin{aligned} |p^k_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{p}^k_{\boldsymbol{\xi}|\boldsymbol{s}}| &\le \left(\frac{p^k \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^k_s, \boldsymbol{\Sigma}^k_{\boldsymbol{ss}}\right)}{f(\boldsymbol{s})} + \frac{\hat{p}^k \mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^k_s, \hat{\boldsymbol{\Sigma}}^k_{ss}\right)}{\hat{f}(\boldsymbol{s})}\right) \left(C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right) \\ \|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{\boldsymbol{\mu}}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\| &\le \left(\frac{\beta}{\alpha} + 1\right)\epsilon_{\boldsymbol{\mu}} + \frac{\alpha+\beta}{\alpha^2}\left(\|\boldsymbol{s}\| + \gamma\right)\epsilon_{\boldsymbol{\Sigma}} \\ \|\boldsymbol{\Sigma}^k_{r|s} - \hat{\boldsymbol{\Sigma}}^k_{r|s}\| &\le \left(\frac{\beta}{\alpha}\right)^2 \epsilon_{\boldsymbol{\Sigma}}. \end{aligned}$$

Hence, Proposition 8 implies that

$$\begin{aligned} &\mathbb{W}_2^2(\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}_{\boldsymbol{\xi}|\boldsymbol{s}}) \\ &\le \left(\left(\frac{\beta}{\alpha} + 1\right)\epsilon_{\boldsymbol{\mu}} + \frac{\alpha+\beta}{\alpha^2}\left(\|\boldsymbol{s}\| + \gamma\right)\epsilon_{\boldsymbol{\Sigma}}\right)^2 + D\left(\frac{\beta}{\alpha}\right)^2 \epsilon_{\boldsymbol{\Sigma}} \\ &\qquad + \left(4\gamma^2 + 2D\beta\right) \sum_{i\in[K]} \left(\frac{p^i \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}^i_s, \boldsymbol{\Sigma}^i_{\boldsymbol{ss}}\right)}{f(\boldsymbol{s})} + \frac{\hat{p}^i \mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}^i_s, \hat{\boldsymbol{\Sigma}}^i_{ss}\right)}{\hat{f}(\boldsymbol{s})}\right) \left(C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right) \\ &\le \left(\left(\frac{\beta}{\alpha} + 1\right)\epsilon_{\boldsymbol{\mu}} + \frac{\alpha+\beta}{\alpha^2}\left(\|\boldsymbol{s}\| + \gamma\right)\epsilon_{\boldsymbol{\Sigma}}\right)^2 + D\left(\frac{\beta}{\alpha}\right)^2 \epsilon_{\boldsymbol{\Sigma}} \\ &\qquad + 2\left(4\gamma^2 + 2D\beta\right)\left(C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right). \end{aligned}$$

This completes the proof. □

Proof of Corollary 1 Since $\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}} \in \mathcal{P}_\varepsilon$, we must have

$$\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \leq \sup_{\mathbb{Q}\in\mathcal{P}_\varepsilon} \mathbb{E}_{\mathbb{Q}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \qquad \forall \boldsymbol{x} \in \mathcal{X}.$$

In particular, substituting the solution $\hat{\boldsymbol{x}}$ into both sides of the inequalities yields the desired guarantee. □

Proof of Proposition 3 By the triangle inequality for the Wasserstein distance, we obtain

$$\mathbb{W}_2(\mathbb{M}^K_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}) \leq \mathbb{W}_2(\mathbb{M}^K_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K}_{\boldsymbol{\xi}|\boldsymbol{s}}) + \mathbb{W}_2(\hat{\mathbb{M}}^{K}_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}).$$

From Theorem 2, we have $\mathbb{W}_2(\mathbb{M}^K_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K}_{\boldsymbol{\xi}|\boldsymbol{s}}) \leq \varepsilon$. Furthermore, from (Delon and Desolneux 2020, Section 4) and (Chen et al. 2018b, Section 3), we obtain that $\mathbb{W}_2(\hat{\mathbb{M}}^{K}_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}})$ is upper bounded by the optimal value of the following linear program:

$$\begin{aligned}
\mathbb{W}_2^2(\hat{\mathbb{M}}^{K}_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}) \leq \overline{\mathbb{W}}_2^2(\hat{\mathbb{M}}^{K}_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}) := \min\ & \sum_{i\in[K]} \sum_{j\in[K']} \pi_{ij} \mathbb{W}_2^2(\hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^i_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^i_{\boldsymbol{\xi}|\boldsymbol{s}}), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^j_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^j_{\boldsymbol{\xi}|\boldsymbol{s}})) \\
\text{s.t.}\ & \boldsymbol{\pi} \in \Delta^K \times \Delta^{K'} \\
& \sum_{i\in[K]} \pi_{ij} = \hat{p}^j \quad \forall j \in [K'] \\
& \sum_{j\in[K']} \pi_{ij} = \hat{p}^i \quad \forall i \in [K],
\end{aligned}$$

where $\mathbb{W}_2^2(\hat{\mathbb{N}}(\boldsymbol{\mu}^i_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^i_{\boldsymbol{\xi}|\boldsymbol{s}}), \hat{\mathbb{N}}(\hat{\boldsymbol{\mu}}^j_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\boldsymbol{\Sigma}}^j_{\boldsymbol{\xi}|\boldsymbol{s}})) = |\hat{\boldsymbol{\mu}}^i_{\boldsymbol{\xi}|\boldsymbol{s}} - \hat{\boldsymbol{\mu}}^j_{\boldsymbol{\xi}|\boldsymbol{s}}\|^2 + \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}^i_{\boldsymbol{\xi}|\boldsymbol{s}} + \hat{\boldsymbol{\Sigma}}^j_{\boldsymbol{\xi}|\boldsymbol{s}} - 2(((\hat{\boldsymbol{\Sigma}}^i_{\boldsymbol{\xi}|\boldsymbol{s}})^{\frac{1}{2}} \hat{\boldsymbol{\Sigma}}^j_{\boldsymbol{\xi}|\boldsymbol{s}} (\hat{\boldsymbol{\Sigma}}^i_{\boldsymbol{\xi}|\boldsymbol{s}})^{\frac{1}{2}})^{\frac{1}{2}}\right)$ (Givens and Shortt 1984, Proposition 7). Since $K \in \mathcal{K}$, we must have $\mathbb{W}_2(\mathbb{M}^K_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}}) \leq \varepsilon + \max_{L\in\mathcal{K}} \overline{\mathbb{W}}_2(\hat{\mathbb{M}}^{L}_{\boldsymbol{\xi}|\boldsymbol{s}}, \hat{\mathbb{M}}^{K'}_{\boldsymbol{\xi}|\boldsymbol{s}})$. The result then follows from optimizing for the best mixture size $K' \in \mathcal{K}$ that minimizes the bound. □

Appendix C: Proofs of Section 4

Proof of Theorem 3 We decompose the error as follows:

$$\begin{aligned}
& \left| \mathbb{E}_{\mathbb{M}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \middle| \tilde{\boldsymbol{s}} = \boldsymbol{s}\right] - \frac{\frac{1}{N}\sum_{n\in[N]} \ell(\boldsymbol{x},\boldsymbol{\xi}_n)\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)}{\hat{f}(\boldsymbol{s})} \right| \\
= & \left| \frac{\int \ell(\boldsymbol{x},\boldsymbol{\xi}) f(\boldsymbol{s},\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{f(\boldsymbol{s})} - \frac{\frac{1}{N}\sum_{n\in[N]} \ell(\boldsymbol{x},\boldsymbol{\xi}_n)\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)}{\hat{f}(\boldsymbol{s})} \right| \\
\leq & \underbrace{\left| \frac{\int \ell(\boldsymbol{x},\boldsymbol{\xi}) f(\boldsymbol{s},\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{f(\boldsymbol{s})} - \frac{\int \ell(\boldsymbol{x},\boldsymbol{\xi}) \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{\hat{f}(\boldsymbol{s})} \right|}_{\text{(a)}} + \underbrace{\left| \frac{\int \ell(\boldsymbol{x},\boldsymbol{\xi}) \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{\hat{f}(\boldsymbol{s})} - \frac{\frac{1}{N}\sum_{n\in[N]} \ell(\boldsymbol{x},\boldsymbol{\xi}_n)\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)}{\hat{f}(\boldsymbol{s})} \right|}_{\text{(b)}}.
\end{aligned}$$

We analyze each term separately:

(a): Using Lemma 2, we have:

$$
\begin{aligned}
&\left| \frac{\int \ell(\boldsymbol{x},\boldsymbol{\xi}) f(\boldsymbol{s},\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{f(\boldsymbol{s})} - \frac{\int \ell(\boldsymbol{x},\boldsymbol{\xi}) \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{\hat{f}(\boldsymbol{s})} \right| \\
&\leq \frac{\left| \int \ell(\boldsymbol{x},\boldsymbol{\xi}) f(\boldsymbol{s},\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} - \int \ell(\boldsymbol{x},\boldsymbol{\xi}) \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} \right|}{f(\boldsymbol{s})} + \frac{|f(\boldsymbol{s}) - \hat{f}(\boldsymbol{s})| \left| \int \ell(\boldsymbol{x},\boldsymbol{\xi}) \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} \right|}{f(\boldsymbol{s}) \hat{f}(\boldsymbol{s})} \\
&\leq \underbrace{\frac{\overline{\ell} \int \left| f(\boldsymbol{s},\boldsymbol{\xi}) - \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi}) \right| \mathrm{d}\boldsymbol{\xi}}{f(\boldsymbol{s})}}_{\text{(i)}} + \underbrace{\frac{|f(\boldsymbol{s}) - \hat{f}(\boldsymbol{s})| \cdot \overline{\ell} \int \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{f(\boldsymbol{s}) \hat{f}(\boldsymbol{s})}}_{\text{(ii)}}.
\end{aligned}
$$

We now bound each term:

(i) We use Proposition 5 to write

$$
\begin{aligned}
|f(\boldsymbol{s},\boldsymbol{\xi}) - \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi})| &\leq |f(\boldsymbol{s},\boldsymbol{\xi}) - \hat{f}(\boldsymbol{s},\boldsymbol{\xi})| + \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) |\hat{f}(\boldsymbol{\xi}) - f(\boldsymbol{\xi})| \\
&\leq f(\boldsymbol{s},\boldsymbol{\xi}) \left(C_{Q+R} \|(\boldsymbol{s},\boldsymbol{\xi})\|^2 + C'_{Q+R} \|(\boldsymbol{s},\boldsymbol{\xi})\| + C''_{Q+R} \right) \\
&\quad + \overline{f} f(\boldsymbol{\xi}) \left(C_R \|\boldsymbol{\xi}\|^2 + C'_R \|\boldsymbol{\xi}\| + C''_R \right),
\end{aligned}
$$

Integrating and applying the assumptions yields:

$$
\begin{aligned}
\text{(i)} &\leq \overline{\ell} \int f(\boldsymbol{\xi}|\boldsymbol{s}) \left(C_{Q+R} \|(\boldsymbol{s},\boldsymbol{\xi})\|^2 + C'_{Q+R} \|(\boldsymbol{s},\boldsymbol{\xi})\| + C''_{Q+R} \right) \mathrm{d}\boldsymbol{\xi} \\
&\quad + \frac{\overline{\ell f}}{\underline{f}} \int f(\boldsymbol{\xi}) \left(C_R \|\boldsymbol{\xi}\|^2 + C'_R \|\boldsymbol{\xi}\| + C''_R \right) \mathrm{d}\boldsymbol{\xi} \\
&\leq \overline{\ell} \mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}} \left[C_{Q+R} \|(\boldsymbol{s},\boldsymbol{\xi})\|^2 + C'_{Q+R} \|(\boldsymbol{s},\boldsymbol{\xi})\| + C''_{Q+R} \right] \\
&\quad + \frac{\overline{\ell f}}{\underline{f}} \mathbb{E}_{\mathbb{M}} \left[C_R \|\boldsymbol{\xi}\|^2 + C'_R \|\boldsymbol{\xi}\| + C''_R \right].
\end{aligned}
$$

Now note:

$$
\begin{aligned}
\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}} \left[\|(\boldsymbol{s},\tilde{\boldsymbol{\xi}})\|^2 \right] &\leq \|\boldsymbol{s}\|^2 + \mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}} \left[\|\tilde{\boldsymbol{\xi}}\|^2 \right] \\
&= \|\boldsymbol{s}\|^2 + \sum_{k \in [K]} p^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathbb{E}_{\mathbb{N}(\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})} \left[\|\tilde{\boldsymbol{\xi}}\|^2 \right] \\
&= \|\boldsymbol{s}\|^2 + \sum_{k \in [K]} p^k_{\boldsymbol{\xi}|\boldsymbol{s}} \mathbb{E}_{\mathbb{N}(\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}, \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}})} \left[\mathrm{tr}(\tilde{\boldsymbol{\xi}} \tilde{\boldsymbol{\xi}}^\top) \right] \\
&= \|\boldsymbol{s}\|^2 + \sum_{k \in [K]} p^k_{\boldsymbol{\xi}|\boldsymbol{s}} \left(\mathrm{tr}\left(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \right) + \|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\|^2 \right).
\end{aligned}
$$

In addition, since

$$
\mathrm{tr}\left(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}|\boldsymbol{s}} \right) = \mathrm{tr}\left(\boldsymbol{\Sigma}^k_{\boldsymbol{\xi}\boldsymbol{\xi}} - \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}\boldsymbol{s}} (\boldsymbol{\Sigma}^k_{\boldsymbol{s}\boldsymbol{s}})^{-1} \boldsymbol{\Sigma}^k_{\boldsymbol{s}\boldsymbol{\xi}} \right) \leq \beta
$$

and

$$
\|\boldsymbol{\mu}^k_{\boldsymbol{\xi}|\boldsymbol{s}}\|^2 = \|\boldsymbol{\mu}^k_{\boldsymbol{\xi}} + \boldsymbol{\Sigma}^k_{\boldsymbol{\xi}\boldsymbol{s}} (\boldsymbol{\Sigma}^k_{\boldsymbol{s}\boldsymbol{s}})^{-1} (\boldsymbol{s} - \boldsymbol{\mu}^k_{\boldsymbol{s}})\|^2
$$

$$\leq \left(\gamma + \frac{\beta}{\alpha}(\|\boldsymbol{s}\| + \gamma)\right)^2,$$

we obtain

$$\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\|(\boldsymbol{s}, \tilde{\boldsymbol{\xi}})\|^2\right] \leq \|\boldsymbol{s}\|^2 + \beta + \left(\gamma + \frac{\beta}{\alpha}(\|\boldsymbol{s}\| + \gamma)\right)^2.$$

Next, by Jensen's inequality

$$\begin{aligned}\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\|(\boldsymbol{s}, \tilde{\boldsymbol{\xi}})\|\right] &\leq \sqrt{\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[\|(\boldsymbol{s}, \tilde{\boldsymbol{\xi}})\|^2\right]} \\ &= \sqrt{\|\boldsymbol{s}\|^2 + \beta + \left(\gamma + \frac{\beta}{\alpha}(\|\boldsymbol{s}\| + \gamma)\right)^2}.\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{E}_{\mathbb{M}_{\boldsymbol{\xi}|\boldsymbol{s}}}\left[C_{Q+R}\|(\boldsymbol{s}, \tilde{\boldsymbol{\xi}})\|^2 + C'_{Q+R}\|(\boldsymbol{s}, \tilde{\boldsymbol{\xi}})\| + C''_{Q+R}\right] \\ &\leq C_{Q+R}\left(\|\boldsymbol{s}\|^2 + \beta + \left(\gamma + \frac{\beta}{\alpha}(\|\boldsymbol{s}\| + \gamma)\right)^2\right) + C'_{Q+R}\sqrt{\|\boldsymbol{s}\|^2 + \beta + \left(\gamma + \frac{\beta}{\alpha}(\|\boldsymbol{s}\| + \gamma)\right)^2} + C''_{Q+R}.\end{aligned}$$

For the marginal expectation:

$$\begin{aligned}&\mathbb{E}_{\mathbb{M}}\left[C_R\|\boldsymbol{\xi}\|^2 + C'_R\|\boldsymbol{\xi}\| + C''_R\right] \\ &\leq C_R\left(\sum_{k\in[K]} p^k\left(\mathrm{tr}(\boldsymbol{\Sigma}^k) + \|\boldsymbol{\mu}\|^2\right)\right) + C'_R\sqrt{\sum_{k\in[K]} p^k\left(\mathrm{tr}(\boldsymbol{\Sigma}^k) + \|\boldsymbol{\mu}\|^2\right)} + C''_R \\ &\leq C_R\left(\beta + \gamma^2\right) + C'_R\sqrt{\beta + \gamma^2} + C''_R.\end{aligned}$$

(ii) Again using Proposition 5, we obtain

$$\frac{\left|f(\boldsymbol{s}) - \hat{f}(\boldsymbol{s})\right|}{f(\boldsymbol{s})} \leq C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q,$$

and so

$$\frac{|f(\boldsymbol{s}) - \hat{f}(\boldsymbol{s})| \cdot \overline{\ell} \int \hat{f}(\boldsymbol{s}|\boldsymbol{\xi}) f(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{f(\boldsymbol{s})\hat{f}(\boldsymbol{s})} \leq \frac{\overline{\ell f}}{\underline{f}}\left(C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right).$$

Combining both bounds gives:

$$\begin{aligned}\text{(a)} \leq\ & \overline{\ell}\Bigg[C_{Q+R}\left(\|\boldsymbol{s}\|^2 + \beta + \left(\gamma + \tfrac{\beta}{\alpha}(\|\boldsymbol{s}\| + \gamma)\right)^2\right) \\ &\qquad + C'_{Q+R}\sqrt{\|\boldsymbol{s}\|^2 + \beta + \left(\gamma + \tfrac{\beta}{\alpha}(\|\boldsymbol{s}\| + \gamma)\right)^2} + C''_{Q+R}\Bigg] \\ &+ \tfrac{\overline{\ell f}}{\underline{f}}\left(C_R\left(\beta + \gamma^2\right) + C'_R\sqrt{\beta + \gamma^2} + C''_R + C_Q\|\boldsymbol{s}\|^2 + C'_Q\|\boldsymbol{s}\| + C''_Q\right).\end{aligned}$$

(b): By Hoeffding's inequality, we have with probability at least $1-\delta$:

$$\left|\int \ell(\boldsymbol{x},\boldsymbol{\xi})\hat{f}(\boldsymbol{s}|\boldsymbol{\xi})f(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} - \frac{1}{N}\sum_{n\in[N]}\ell(\boldsymbol{x},\boldsymbol{\xi}_n)\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)\right| \leq \overline{\ell f}\sqrt{\frac{1}{2N}\log\left(\frac{4}{\delta}\right)}.$$

Thus,

$$\left|\frac{\int \ell(\boldsymbol{x},\boldsymbol{\xi})\hat{f}(\boldsymbol{s}|\boldsymbol{\xi})f(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}}{\hat{f}(\boldsymbol{s})} - \frac{\frac{1}{N}\sum_{n\in[N]}\ell(\boldsymbol{x},\boldsymbol{\xi}_n)\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)}{\hat{f}(\boldsymbol{s})}\right| \leq \frac{\overline{\ell f}}{\underline{f}}\sqrt{\frac{1}{2N}\log\left(\frac{4}{\delta}\right)}.$$

Finally, combining (a), (b), and applyingthe union bound with Assumption (D), we have

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{M}}\left[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\middle|\tilde{\boldsymbol{s}}=\boldsymbol{s}\right] - \frac{1}{\hat{f}(\boldsymbol{s})N}\sum_{n\in[N]}\ell(\boldsymbol{x},\boldsymbol{\xi}_n)\hat{f}(\boldsymbol{s}|\boldsymbol{\xi}_n)\right| \\
\leq\ & \overline{\ell}\left[C_{Q+R}\left(\|\boldsymbol{s}\|^2+\beta+\left(\gamma+\tfrac{\beta}{\alpha}(\|\boldsymbol{s}\|+\gamma)\right)^2\right) + C'_{Q+R}\sqrt{\|\boldsymbol{s}\|^2+\beta+\left(\gamma+\tfrac{\beta}{\alpha}(\|\boldsymbol{s}\|+\gamma)\right)^2} + C''_{Q+R}\right] \\
&+\tfrac{\overline{\ell f}}{\underline{f}}\left(C_R\left(\beta+\gamma^2\right)+C'_R\sqrt{\beta+\gamma^2}+C''_R+C_Q\|\boldsymbol{s}\|^2+C'_Q\|\boldsymbol{s}\|+C''_Q\right) \\
&+\tfrac{\overline{\ell f}}{\underline{f}}\sqrt{\tfrac{1}{2N}\log\left(\tfrac{4}{\delta}\right)}.
\end{aligned}$$

holds with probability at least 1-2δ. Setting $2\delta \to \delta$ completes the proof. □

The proof of Theorem 4 relies on the following preliminary result.

LEMMA 7. *Let $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^d$ be a random vector drawn from $\mathbb{M} = \sum_{k\in[K]} p^k \mathbb{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$. Then, the centered vector $\tilde{\boldsymbol{\xi}} - \mathbb{E}[\tilde{\boldsymbol{\xi}}]$ is sub-Gaussian, and its squared sub-Gaussian parameter σ^2 satisfies:*

$$\sigma^2 \leq \max_{j\in[K]}\left\|\boldsymbol{\Sigma}^j\right\|_2 + \sum_{k=1}^{K} p^k\left\|\boldsymbol{\mu}^k - \bar{\boldsymbol{\mu}}\right\|_2^2 \tag{24}$$

where $\bar{\boldsymbol{\mu}} = \mathbb{E}[\tilde{\boldsymbol{\xi}}] = \sum_{k=1}^{K} p^k \boldsymbol{\mu}^k$ is the global mean of the mixture.

Proof of Lemma 7 To establish the sub-Gaussian property, we bound the moment generating function (MGF) of the projection $\boldsymbol{u}^\top(\tilde{\boldsymbol{\xi}} - \bar{\boldsymbol{\mu}})$ for any unit vector $\boldsymbol{u} \in \mathbb{R}^d$. By the law of total expectation, the MGF can be written as a convex combination of the component MGFs:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(t\boldsymbol{u}^\top(\tilde{\boldsymbol{\xi}}-\bar{\boldsymbol{\mu}})\right)\right] &= \sum_{k=1}^{K} p^k \mathbb{E}\left[\exp\left(t\boldsymbol{u}^\top(\tilde{\boldsymbol{\xi}}^k-\bar{\boldsymbol{\mu}})\right)\right] \\
&= \sum_{k=1}^{K} p^k \exp\left(t\boldsymbol{u}^\top(\boldsymbol{\mu}^k-\bar{\boldsymbol{\mu}}) + \frac{t^2}{2}(\boldsymbol{u}^\top\boldsymbol{\Sigma}^k\boldsymbol{u})\right),
\end{aligned}$$

where $\tilde{\boldsymbol{\xi}}^k \sim \mathcal{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$. We can bound the quadratic form $\boldsymbol{u}^\top\boldsymbol{\Sigma}^k\boldsymbol{u} \leq \left\|\boldsymbol{\Sigma}^k\right\|_2$, where $\|\cdot\|_2$ is the spectral norm. This allows us to factor out a uniform bound for the variance terms:

$$\mathbb{E}\left[\exp\left(t\boldsymbol{u}^\top(\tilde{\boldsymbol{\xi}}-\bar{\boldsymbol{\mu}})\right)\right] \leq \exp\left(\frac{t^2}{2}\max_{j\in[K]}\left\|\boldsymbol{\Sigma}^j\right\|_2\right)\left(\sum_{k=1}^{K} p^k \exp\left(t\boldsymbol{u}^\top(\boldsymbol{\mu}^k-\bar{\boldsymbol{\mu}})\right)\right). \tag{25}$$

The remaining sum is the MGF of $\boldsymbol{u}^\top(\tilde{\boldsymbol{\zeta}} - \bar{\boldsymbol{\mu}})$, where $\tilde{\boldsymbol{\zeta}}$ is an auxiliary random vector that takes value $\boldsymbol{\mu}^k$ with probability p^k. As $\boldsymbol{u}^\top(\tilde{\boldsymbol{\zeta}} - \bar{\boldsymbol{\mu}})$ is a zero-mean, bounded random variable, its MGF is bounded by the MGF of a Gaussian with the same variance. By applying Jensen's inequality to the convex exponential function, we have:

$$\begin{aligned}
\sum_{k=1}^{K} p^k \exp\left(t\boldsymbol{u}^\top(\boldsymbol{\mu}^k - \bar{\boldsymbol{\mu}})\right) &\leq \mathbb{E}\left[\exp\left(\frac{t^2}{2}\mathbb{V}\left(\boldsymbol{u}^\top(\tilde{\boldsymbol{\zeta}} - \bar{\boldsymbol{\mu}})\right)\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{t^2}{2}\left(\boldsymbol{u}^\top(\tilde{\boldsymbol{\zeta}} - \bar{\boldsymbol{\mu}})\right)^2\right)\right] \\
&\leq \exp\left(\frac{t^2}{2}\mathbb{E}\left[\left(\boldsymbol{u}^\top(\tilde{\boldsymbol{\zeta}} - \bar{\boldsymbol{\mu}})\right)^2\right]\right) \\
&\leq \exp\left(\frac{t^2}{2}\mathbb{E}\left[\left\|\tilde{\boldsymbol{\zeta}} - \bar{\boldsymbol{\mu}}\right\|_2^2\right]\right) \\
&= \exp\left(\frac{t^2}{2}\sum_{k=1}^{K} p^k \left\|\boldsymbol{\mu}^k - \bar{\boldsymbol{\mu}}\right\|_2^2\right),
\end{aligned}$$

where the first inequality is based on the definition of sub-Gaussian random variables (Rigollet and Hütter 2023, Definition 1.2), and the second equality comes from the fact that $\boldsymbol{u}^\top(\tilde{\boldsymbol{\zeta}} - \bar{\boldsymbol{\mu}})$ has a mean of zero. Substituting this back into inequality (25) yields:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(t\boldsymbol{u}^\top(\tilde{\boldsymbol{\xi}} - \bar{\boldsymbol{\mu}})\right)\right] &\leq \exp\left(\frac{t^2}{2}\max_j \left\|\boldsymbol{\Sigma}^j\right\|_2\right)\exp\left(\frac{t^2}{2}\sum_{k=1}^{K} p^k \left\|\boldsymbol{\mu}^k - \bar{\boldsymbol{\mu}}\right\|_2^2\right) \\
&= \exp\left(\frac{t^2}{2}\left(\max_j \left\|\boldsymbol{\Sigma}^j\right\|_2 + \sum_{k=1}^{K} p^k \left\|\boldsymbol{\mu}^k - \bar{\boldsymbol{\mu}}\right\|_2^2\right)\right).
\end{aligned}$$

Since this bound holds for any unit vector $\boldsymbol{u}$, the result follows from the definition of a sub-Gaussian random vector. □

We now proceed to show the derivation of Theorem 4.

Proof of Theorem 4 We first establish a high probability bound on the 2-norm of the sample points $\{\boldsymbol{\xi}_{t,n}\}_{t\in[T],n\in[N]}$. Since $\boldsymbol{\xi}_{t,n} \in \mathbb{R}^R$ sampled from a sub-Gaussian distribution with variance proxy $\sigma^2 \leq \max_{j\in[K]} \left\|\boldsymbol{\Sigma}^j\right\|_2 + \sum_{k=1}^{K} p^k \left\|\boldsymbol{\mu}^k - \bar{\boldsymbol{\mu}}\right\|_2^2$, then by (Rigollet and Hütter 2023, Theorem 1.19), we have with probability at least $1-\delta$,

$$\|\boldsymbol{\xi}_{t,n}\| \leq \bar{\xi} := 4\sigma\sqrt{R} + 2\sigma\sqrt{2\log\tfrac{NT}{\delta}} + \gamma \quad \forall n \in [N] \ \forall t \in [T]. \tag{26}$$

Next, we ensure that the bound in (12) holds uniformly for all $\boldsymbol{x} \in \mathcal{X}_t(\cdot)$. To simplify the exposition, we define

$$\begin{aligned}
\tau := \bar{\ell}\Bigg[& C_{Q+R}\left(\bar{\xi}^2 + \beta + \left(\gamma + \tfrac{\beta}{\alpha}(\bar{\xi}+\gamma)\right)^2\right) + C'_{Q+R}\sqrt{\bar{\xi}^2 + \beta + \left(\gamma + \tfrac{\beta}{\alpha}(\bar{\xi}+\gamma)\right)^2} + C''_{Q+R}\Bigg] \\
& + \tfrac{\overline{\ell f}}{\underline{f}}\left(C_R\left(\beta+\gamma^2\right) + C'_R\sqrt{\beta+\gamma^2} + C''_R + C_Q\bar{\xi}^2 + C'_Q\bar{\xi} + C''_Q\right)
\end{aligned}$$

Note that by construction, τ upper bounds the error expression in (12) for all sample points satisfying (26). Thus, by (12), we have

$$\left|\mathbb{E}\left[V_{t+1}(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{t+1})\middle|\tilde{\boldsymbol{\xi}}_t=\boldsymbol{\xi}_t\right]-\hat{\mathbb{E}}\left[V_{t+1}(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{t+1})\middle|\tilde{\boldsymbol{\xi}}_t=\boldsymbol{\xi}_t\right]\right|\le\tau+\frac{\overline{\ell}\,\overline{f}^2}{\underline{f}^2}\sqrt{\frac{2}{N}\log\left(\frac{8}{\delta}\right)}$$

with probability at least $1-\delta$. Using a covering number argument, we obtain the uniform bound

$$\left|\mathbb{E}\left[V_{t+1}(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{t+1})\middle|\tilde{\boldsymbol{\xi}}_t=\boldsymbol{\xi}_t\right]-\hat{\mathbb{E}}\left[V_{t+1}(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{t+1})\middle|\tilde{\boldsymbol{\xi}}_t=\boldsymbol{\xi}_t\right]\right|\le\tau+\frac{\overline{\ell}\,\overline{f}^2}{\underline{f}^2}\sqrt{\frac{2}{N}\log\left(\frac{\mathcal{O}(\overline{D}/\eta)^R}{\delta}\right)}+2L\eta$$

for all $\boldsymbol{x}_t\in\mathcal{X}_t(\boldsymbol{x}_{t-1},\boldsymbol{\xi}_t)$ with probability at least $1-\delta$. Following the recursive induction argument in the proof of (Park et al. 2022, Theorem 1), we propagate the stagewise approximation errors backward from $t=T$ to $t=1$, yielding:

$$\begin{aligned}&\boldsymbol{\ell}(\widehat{\boldsymbol{x}}_1^N,\boldsymbol{\xi}_1)+\mathcal{V}_2(\widehat{\boldsymbol{x}}_1^N,\boldsymbol{\xi}_1)-\left(\min_{\boldsymbol{x}_1\in\mathcal{X}_1(\boldsymbol{x}_0,\boldsymbol{\xi}_1)}\ell(\boldsymbol{x}_1,\boldsymbol{\xi}_1)+\mathcal{V}_2(\boldsymbol{x}_1,\boldsymbol{\xi}_1)\right)\\&\qquad\le 2(T-1)\tau+4(T-1)L\eta+2\sum_{t=2}^{T}\frac{\overline{\ell}\,\overline{f}^2}{\underline{f}^2}\sqrt{\frac{2}{N}\log\left(\frac{O(1)N^{t-2}(\overline{D}/\eta)^{R(t-1)}}{\delta}\right)}\end{aligned}$$

with probability at least $1-(T-1)\delta$. Finally, applying the union bound with (26) and Assumption (D), we conclude that the result holds with probability at least $1-(T+1)\delta$. Finally, setting $(T+1)\delta\to\delta$ completes the proof. □

Appendix D: Details of Experiments

D.1. Inventory Management

For the contextual newsvendor problem, the distributionally robust optimization model derived from Proposition 2 is formulated as the following SOCP:

$$\begin{aligned}\inf\ &\lambda\varepsilon^2+\frac{1}{M}\sum_{m\in[M]}s_m\\ \text{s.t. }&\lambda\in\mathbb{R}_+,\boldsymbol{s}\in\mathbb{R}^Q,\theta\in\mathbb{R}_+,q\in\mathbb{R}_+\\ &\left\|\begin{bmatrix}2\lambda\xi_m+\theta-h\\ \lambda\xi_m^2-hq+s_m-\lambda\end{bmatrix}\right\|_2\le\lambda\xi_m^2-hq+s_m+\lambda && \forall m\in[M]\\ &\left\|\begin{bmatrix}2\lambda\xi_m+\theta+b\\ \lambda\xi_m^2+bq+s_m-\lambda\end{bmatrix}\right\|_2\le\lambda\xi_m^2+bq+s_m+\lambda && \forall m\in[M]\\ &\lambda\xi_m^2-hq+s_m\ge 0 && \forall m\in[M]\\ &\lambda\xi_m^2+bq+s_m\ge 0 && \forall m\in[M]\end{aligned}$$

The implementation of our framework requires tuning of several hyperparameters. The number of mixture components K in the GMM is selected using the Akaike Information Criterion

(AIC)(Akaike 2025), which is particularly well-suited for predictive tasks with moderate sample sizes(Chakrabarti and Ghosh 2011). Compared to alternatives such as the Bayesian Information Criterion (BIC), AIC favors models with better predictive accuracy in finite-sample settings. Once K is determined, we estimate the GMM parameters.

To determine the Wasserstein radius hyperparameter ϵ, we adopt a hold-out validation approach. Specifically, we reserve 10% of the training data as a validation set. For a given candidate value of ϵ from the set $\{0.01, 0.05, 0.1, 0.5, 1\}$, the model is trained on the remaining 90% of the data. The value of ϵ that yields the minimum average validation loss is then selected for the final model.

For the GMM-NF model designed to handle general distributions, we employ an Autoregressive Rational Quadratic Spline (ARQS) flow, which is known for its strong expressiveness and ability to approximate complex, high-dimensional distributions (Durkan et al. 2019). The number of components for the latent GMM is also selected via AIC. To prevent overfitting, particularly when the training data is sparse, we adopt a lightweight ARQS architecture consisting of 64 hidden nodes, a single hidden layer, a single block size, and 8 spline bins. The flow is trained for a maximum of 200 epochs, and we implement early stopping based on the validation loss, calculated on a random 20% split of the training data. Training is halted if the validation loss fails to improve for 30 consecutive epochs. This design balances model flexibility and generalization, ensuring stable training even with limited in-sample data.

D.2. Portfolio Optimization

First, we detail the preprocessing of the side information. We employ a Gaussian kernel bandwidth selection procedure for the contextual factors. Using data from the initial training period (January 1, 2017, to December 31, 2020), we treat each of the five financial indicators as a predictor variable in a Nadaraya-Watson regression, with the unweighted average stock return across the asset universe serving as the response variable. The optimal Gaussian kernel bandwidth for each predictor is then determined via least-squares cross-validation. Subsequently, the raw values of each side information index are scaled by dividing them by their corresponding optimal bandwidth. This scaling ensures that each factor contributes appropriately to the model without being dominated by differences in their native scales.

By applying the 2-Wasserstein robust formulation from Proposition 2, our model can be formulated as the following SOCP:

$$
\begin{aligned}
\inf_{\lambda,\boldsymbol{\gamma},\beta} \quad & \lambda\varepsilon^2 + \frac{1}{M}\sum_{m\in[M]} \boldsymbol{\gamma}_i \\
\text{s.t.} \quad & \lambda \in \mathbb{R}_+, \quad \boldsymbol{x}\in\mathbb{R}_+^D, \quad \boldsymbol{\gamma}\in\mathbb{R}^{M\times D}, \quad \beta\in\mathbb{R} \\
& \left\| \begin{bmatrix} 2\lambda - \left(\frac{1}{\tau}+\eta\right)\boldsymbol{x} \\ \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \boldsymbol{\gamma}_m + \dfrac{\beta}{\tau} - \beta - \lambda \end{bmatrix} \right\| \le \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \boldsymbol{\gamma}_m + \frac{\beta}{\tau} - \beta + \lambda \quad && \forall m\in[M] \\
& \left\| \begin{bmatrix} 2\lambda\boldsymbol{\xi}_m - \eta\boldsymbol{x} \\ \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \boldsymbol{\gamma}_m + \dfrac{\beta}{\tau} - \beta - \lambda \end{bmatrix} \right\| \le \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \boldsymbol{\gamma}_m - \beta + \lambda \quad && \forall m\in[M] \\
& \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \boldsymbol{\gamma}_m \ge -\frac{\beta}{\tau} + \beta && \forall m\in[M] \\
& \lambda\boldsymbol{\xi}_m^\top\boldsymbol{\xi}_m + \boldsymbol{\gamma}_m \ge \beta && \forall m\in[M] \\
& \sum_{d\in[D]} x_d = 1
\end{aligned}
$$

For training the Normalizing Flow, we use an Autoregressive Rational Quadratic Spline (ARQS) flow to model the latent data distribution, leveraging its strong expressiveness for complex, high-dimensional data (Durkan et al. 2019). The number of clusters for the latent GMM is determined by AIC. To mitigate the risk of overfitting in a data-sparse, high-dimensional setting, we adopt a lightweight ARQS architecture with 32 hidden nodes, 1 hidden layer, 1 block, and 8 spline bins for low, medium dimensions, and 64 hidden nodes, 1 hidden layer, 1 block, and 8 spline bins for high dimensions. Training is conducted for up to 200 epochs with an early stopping mechanism: on a random 80/20 split of the training data, if the validation loss does not improve for 30 consecutive epochs, the training is halted. This setup ensures a balance between model flexibility and generalization.

Finally, the hyperparameter tuning for the Wasserstein radius ϵ is performed using a cross-validation procedure. We utilize the designated validation period (January 1, 2019, to December 31, 2020) and randomly sample 50 days from this horizon. For each of these 50 validation days, we construct a corresponding training set using a rolling window of the preceding two years of data. The model is trained for each candidate value of ϵ from the set $\{0.01, 0.05, 0.09, 0.1, 0.5, 0.9\}$. The validation loss is then computed based on the next day's return. The optimal ϵ is chosen as the one that achieves the lowest average validation loss across these 50 days, ensuring our final model is well-calibrated for out-of-sample performance.

D.3. Wind Energy Multi-Stage Optimization

We adopt the multistage framework of Park et al. (2022). Consider three storage units indexed by $\ell \in [3]$, each characterized by a capacity $\bar{s}^\ell$, leakage rate γ^ℓ, and charging/discharging efficiencies γ_c^ℓ and γ_d^ℓ. We let $\boldsymbol{e}_{t+1}^s, \boldsymbol{e}_{t+1}^u, \boldsymbol{e}_{t+1}^d \in \mathbb{R}_+^{24}$ denote, respectively, the wind energy allocated to commitments, the unmet demand, and the curtailled energy. Furthermore, $\boldsymbol{e}_{t+1}^{+,\ell}$ and $\boldsymbol{e}_{t+1}^{-,\ell}$ represent the charging and discharging flows for storage ℓ, and $\boldsymbol{s}_{t+1}^\ell$ its state of charge.

Under these definitions, and taking the random parameter as wind generation $\xi_t = w_t$, the dynamic programming recursion is written as

$$
\begin{aligned}
Q_t(\boldsymbol{s}_t, \boldsymbol{\xi}_t) = \max\ & \boldsymbol{p}_t^\top \boldsymbol{u}_t - 2\boldsymbol{p}_t^\top \mathbb{E}[\boldsymbol{e}_{t+1}^u | \boldsymbol{\xi}_t] + \mathbb{E}[Q_{t+1}(\boldsymbol{s}_{t+1}, \tilde{\boldsymbol{\xi}}_{t+1} | \boldsymbol{\xi}_t)] \\
\text{s.t. } & \boldsymbol{u}_t, \boldsymbol{e}_{t+1}^{\{u,s,d\}} \in \mathbb{R}_+^{24},\ \boldsymbol{e}_{t+1}^{\{+,-\},\ell}, \boldsymbol{s}_{t+1}^\ell \in \mathbb{R}_+^{24} && \forall \ell \in [3] \\
& w_{t+1,h} = e_{t+1,h}^s + e_{t+1,h}^{+,1} + e_{t+1,h}^{+,2} + e_{t+1,h}^{+,3} + e_{t+1,h}^d && \forall h \in [24] \\
& u_{t,h} = e_{t+1,h}^s + e_{t+1,h}^{-,1} + e_{t+1,h}^{-,2} + e_{t+1,h}^{-,3} + e_{t+1,h}^u && \forall h \in [24] \\
& s_{t+1,h}^\ell = \gamma_{t+1,h-1}^\ell + \gamma_c^\ell e_{t+1,h}^{+,\ell} - \frac{1}{\gamma_d^\ell} e_{t+1,h}^{-,\ell} && \forall h \in [24], \forall \ell \in [3] \\
& s_{t+1,h}^\ell \le \bar{s}^\ell && \forall h \in [24], \forall \ell \in [3]
\end{aligned}
$$

To model temporal dependence in wind energy generation, we build a Gaussian mixture over consecutive daily generated wind energy vectors. Let $\boldsymbol{\xi}_t \in \mathbb{R}_+^{24}$ denote the 24-hour wind vector on day t. This yields closed-form *density–ratio* transition weights

$$
w_{ij} \propto \frac{f(\boldsymbol{\xi}_t^i | \boldsymbol{\xi}_{t+1}^j)}{f(\boldsymbol{\xi}_t^i)}
$$

, where i and j are the scenario at each stage t and $t+1$. For robustness, we scale the ambiguity radius η by the empirical 90th percentile and test different values of the scaled parameter η_{scaled} to identify the most effective setting. In our experiments, we use $\eta = 100$ for the GMM-based model with $K = 4$ components in both Ohio and North Carolina, $\eta = 2000$ for the Nadaraya–Watson scheme in Ohio, and $\eta = 1000$ for the Nadaraya–Watson scheme in North Carolina.

We run 10 iterations of the forward–backward procedure. The initial storage capacities are set to $\{5000, 2000, 1000\}$, with leakage rates $\gamma^\ell = \{0.98, 0.99, 0.995\}$, charging efficiencies $\gamma_c^\ell = \{0.8, 0.9, 1.0\}$, and discharging efficiencies $\gamma_d^\ell = \{0.8, 0.9, 1.0\}$. All storage units are initialized at full capacity.